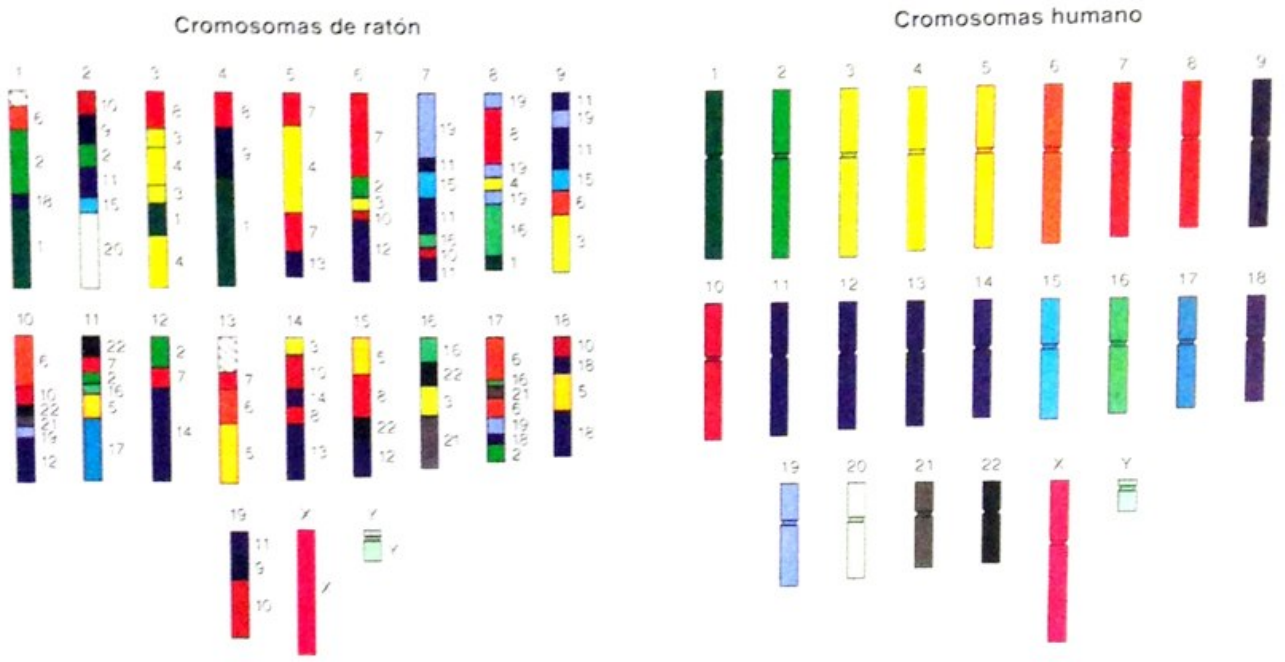


Secuencias genómicas y cantidad de genes

Los cromosomas humano y de ratón. Los genomas de ambas especies experimentaron reordenamientos cromosómicos a partir del genoma de su ancestro común más reciente. Los colores y números que corresponden a los cromosomas del ratón indican los cromosomas humanos que contienen segmentos homólogos. Fotografía cortesía de *U. S. Department of Energy*. Con autorización de Lisa J. Stubbs, *University of Illinois at Urbana–Champagne*.

ÍNDICE

FIGURA 5.1 El número mínimo de genes necesario para cualquier tipo de organismo se incrementa con su complejidad. Fotografía de una bacteria intracelular, cortesía de Gregory P. Henderson y Grant J. Jensen, California *Institute of Technology*. Fotografía de una bacteria de vida libre, cortesía de Kark O. Stetter, *Universität Regensburg*. Fotografía de un eucarionte unicelular, cortesía de Eishi Noguchi, Drexel *University College of Medicine*. Fotografía de un eucarionte multicelular, cortesía de Carolyn B. Marks y David H. Hall, *Albert Einstein College of Medicine*, Bronx, Nueva York. Fotografía de una planta superior, cortesía de Keith Weller/USDA. Fotografía de mamífero © Photodisc.

5.1 Introducción

Desde que se secuenciaron los primeros genomas de organismos, en 1995, se mejoró en gran medida tanto la velocidad como el alcance de la secuenciación. Los primeros genomas secuenciados eran genomas bacterianos pequeños, < 2 Mb. Para el 2002, se había secuenciado el genoma humano, de > 3.000 Mb. En la actualidad, se han secuenciado los genomas de una amplia variedad de organismos, que incluyen bacterias, *archaeas*, levaduras y otros eucariontes unicelulares, plantas, y animales como gusanos, moscas y mamíferos.

Tal vez la información más importante que ha proporcionado una secuencia de genoma haya sido la cantidad de genes que posee (véase la *Sección 4.1, Introducción* para un análisis acerca de las dificultades de definir un gen; para los propósitos de este capítulo, el término "gen" se refiere a una secuencia de DNA que se transcribe a mRNA, tRNA o rRNA). *Mycoplasma genitalium*, una bacteria parasitaria de vida libre, tiene el genoma más pequeño conocido de cualquier organismo, con sólo ~470 genes. Los genomas de bacterias de vida libre tienen desde 1.700 hasta 7.500 genes. Los genomas de *archaea* tienen una variación similar. Los genomas de eucariontes unicelulares comienzan con alrededor de 5.300 genes. Los gusanos y las moscas tienen apenas 18.500 y 13.500 genes, respectivamente, pero la cantidad se eleva a sólo ~25.000 para los genomas humano y de ratón.

La **FIGURA 5.1** resume la cantidad mínima de genes que se encuentran en seis grupos de organismos. Una célula requiere ~500 genes, una célula de vida libre requiere ~1.500 genes, una célula con núcleo requiere > 5.000, un organismo multicelular requiere >10.000 genes, y un organismo con sistema nervioso requiere > 13.000 genes. Muchas especies pueden tener más de la cantidad mínima de genes requeridos, de manera tal que el número de genes puede variar ampliamente, incluso entre especies que están estrechamente relacionadas.

Dentro de los eucariontes unicelulares y de las bacterias, la mayoría de los genes son únicos. Sin embargo, dentro de los genomas eucariontes multicelulares, algunos genes se disponen en familias de miembros relacionados. Por supuesto, algunos genes son únicos (lo que significa que la familia tiene sólo un miembro), pero muchos pertenecen a familias con diez o más miembros. La cantidad de familias diferentes puede ser un mejor indicador de la complejidad general del organismo que la cantidad de genes.

Parte de la información más reveladora proviene de la comparación de las secuencias de genomas. Con las secuencias que se disponen en la actualidad de los genomas de humano y de chimpancé, es posible comenzar a responder algunas de las preguntas acerca de qué convierte a los seres humanos en únicos.

5.2 El número de genes procariontes varía por encima de un orden de magnitud

Actualmente, los esfuerzos a gran escala han conducido a la secuenciación de muchos genomas. En la **FIGURA 5.2**, se resume la variación de los tamaños de los genomas más conocidos. Éstos se extienden desde los $0{,}6 \times 10^6$ pb de un micoplasma a los $3{,}3 \times 10^9$ pb del genoma humano, e incluyen diversos animales experimentales importantes, como las levaduras, la mosca de la fruta y un gusano nemátodo.

Todas las secuencias de los genomas de procariontes demuestran que virtualmente todo el DNA (en general, del 85% al 90%) codifica RNA o polipéptidos. La **FIGURA 5.3** muestra que el rango de los tamaños de los genomas el alrededor de un orden de magnitud, y que el tamaño del genoma es proporcional a la cantidad de genes. El gen típico promedio tiene alrededor de 1.000 pb de longitud.

Todos los procariontes con tamaños de genomas inferiores a 1,5 Mb son parásitos –éstos pueden vivir dentro de un huésped eucarionte que les proporciona moléculas pequeñas. Los tamaños de sus genomas sugieren la cantidad mínima de funciones requeridas para un organismo celular. Todas las clases de genes se encuentran en cantidad reducida en comparación con las de los procariontes con genomas más grandes; pero la reducción más significativa está en los loci, que codifican las enzimas concernientes a las funciones metabólicas (que son, en gran medida, proporcionadas por la célula huésped), y a la regulación de la expresión génica. *Mycoplasma genitalium* tiene el genoma más pequeño, con ~470 genes.

Las *archaeas* tienen propiedades biológicas que son intermedias entre las de otros procariontes y las de eucariontes, pero sus tamaños de genomas y las cantidades de genes se encuentran dentro del mismo rango que los de las bacterias. Sus tamaños de genomas varían desde 1,5 a 3 Mb, correspondiendo a 1.500 a 2.700 genes. *Methanococcus jannaschii* es una especie productora de metano que vive bajo presión y temperatura elevadas. La cantidad total de sus genes es similar a la de *Haemophilus influenzae*, pero pocos de ellos pueden identificarse sobre la base de la comparación con los genes conocidos de otros organismos. Su aparato para la expresión génica se asemeja al de los eucariontes más que al de los procariontes, pero su aparato para la división celular se asemeja más al de los procariontes.

Los genomas de *archaea* y los de las bacterias de vida libre más pequeñas sugieren las cantidades mínimas de genes necesarios para constituir una célula capaz de funcionar de forma independiente en su ambiente. El genoma de *archaea* más pequeño tiene ~1.500 genes. La bacteria no parasitaria de vida libre con el genoma más pequeño conocido es el termófilo *Aquifex aeolicus*, con un genoma de 1,5 Mb y 1.512 genes. Una "típica" bacteria gramnegativa, *H. influenzae*, tiene 1.743 genes, cada uno de los cuales es de ~900 pb. Por lo tanto, se puede concluir que un organismo exclusivamente de vida libre necesita ~1.500 genes.

Los tamaños de genomas procariontes se extienden por encima de un orden de magnitud, desde 0,6 Mb hasta < 8 Mb. Como se esperaba, los genomas más grandes tienen más genes. Los procariontes con los genomas más grandes –*Sinorhizobium meliloti* y *Mesorhizobium loti*– son bacterias fijadoras del nitrógeno que viven sobre las raíces de las plantas. Sus tamaños de genomas (~7 Mb) y la cantidad total de genes (> 7.500) son similares a los de la levadura.

Especies	Genomas (Mb)	Genes	Loci letal
Mycoplasma genitalium	0.58	470	~300
Rickettsia prowazekii	1.11	834	
Haemophilus influenzae	1.83	1.743	
Methanococcus jannaschi	1.66	1.738	
B. subtilis	4.2	4.100	
E. coli	4.6	4.288	1.800
S. cerevisiae	13.5	6.034	1.090
S. pombe	12.5	4.929	
A. thaliana	119	25.498	
O. sativa (arroz)	466	~30.000	
D. melanogaster	165	13.601	3.100
C. elegans	97	18.424	
H. sapiens	3.300	~25.000	

FIGURA 5.2 A partir de la secuencia completa, se conocen los tamaños de los genomas y los números de genes para diversos organismos. Los loci letales se estiman a partir de la información genética.

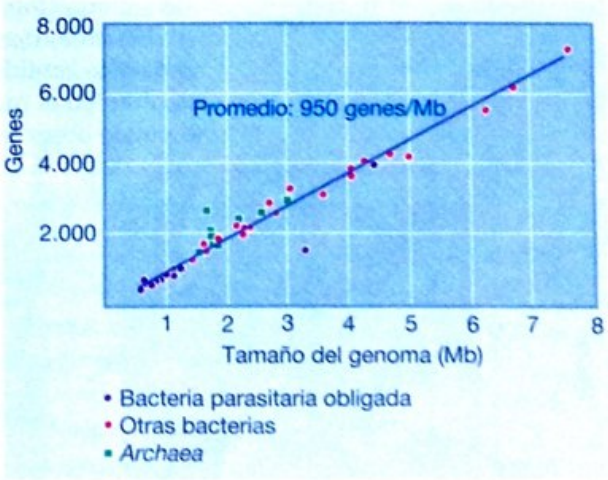


FIGURA 5.3 El número de genes en los genomas bacterianos y de *archaea* es proporcional al tamaño del genoma.

El tamaño del genoma de *E. coli* se encuentra en el medio del intervalo para los procariontes. La cepa de laboratorio común tiene 4.288 genes, con una longitud promedio de ~950 pb y una separación promedio entre genes de 118 pb. No obstante, pueden existir diferencias significativas entre las cepas. Los extremos conocidos entre las cepas de *E. coli* se encuentran desde 4,6 Mb, con 4.249 genes, hasta 5,5 Mb, con 5.361 genes.

Aún no se conocen las funciones de todos estos genes. En la mayoría de estos genomas, ~60% de los genes puede identificarse sobre la base de la homología con genes conocidos de otras especies. Estos genes se encuentran de manera casi equitativa dentro de clases cuyos productos se vinculan con funciones relacionadas con el metabolismo, con la estructura celular o el transporte de componentes, y con la expresión génica y su regulación. Virtualmente, en cada genoma, a más del 25% de los genes no se les puede asignar función. Muchos de estos genes se pueden encontrar en organismos relacionados, lo que implica que tienen una función conservada.

Existe algún énfasis en la secuenciación de los genomas de bacterias patógenas, dada su importancia médica. Una señal importante acerca de la naturaleza de la patogenia ha sido proporcionada por la demostración de que las **islas de patogenicidad** son una característica típica de sus genomas. Éstas son regiones grandes, ~10 a 200 kb, que están presentes en los genomas de especies patógenas, pero ausente en los genomas de variantes no patógenas de la misma especie o de una relacionada. Su contenido de G–C suele ser diferente al del resto del genoma, y es probable que estas islas de patogenia migren entre las bacterias mediante un proceso de **transferencia horizontal**. Por ejemplo, la bacteria que causa el carbunco (*Bacillus anthracis*) tiene dos plásmidos grandes (DNA extracromosómico), uno de los cuales tiene una isla de patogenia que incluye el gen que codifica la toxina del carbunco.

- **islas de patogenicidad:** Segmentos de DNA presentes en genomas bacterianos patogénicos, pero ausentes en bacterias no patogénicas relacionadas.
- **transferencia horizontal:** Transferencia de DNA de una célula a otra por un proceso que no es el de la división celular, como la conjugación bacteriana.

CONCEPTOS CLAVE

- La cantidad mínima de genes para un procarionte parasitario es alrededor de 500; para un procarionte no parasitario de vida libre, alrededor de 1.500.

REVISIÓN DE CONCEPTOS

¿Por qué algunos procariontes tienen un número mayor de genes mientras que otros proliferan con mucha menos cantidad?

5.3 Se conoce el número total de genes para diversos eucariontes

No bien se observan los genomas de eucariontes, la relación entre el tamaño del genoma y la cantidad de genes se debilita. Los genomas de eucariontes unicelulares se encuentran en el mismo rango de tamaño que los genomas bacterianos más grandes. Como se puede observar en la **FIGURA 5.4**, los eucariontes multicelulares tienen más genes, pero su cantidad no se correlaciona con el tamaño del genoma.

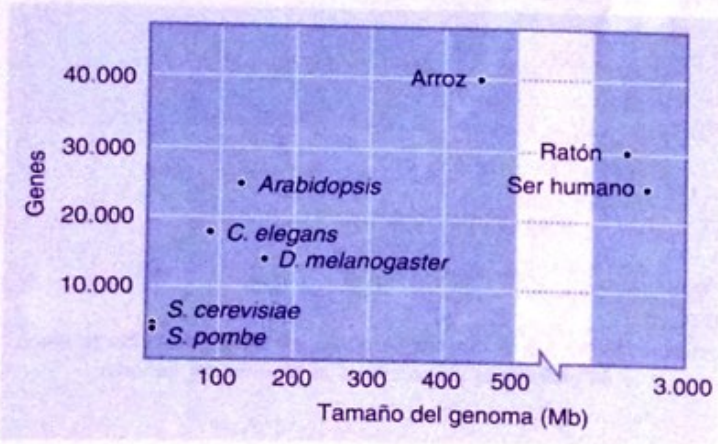


FIGURA 5.4 El número de genes en un eucarionte varía desde 6.000 a 40.000, pero no se correlaciona con el tamaño del genoma o la complejidad del organismo.

La información más amplia para eucariontes unicelulares se dispone a partir de las secuencias de los genomas de las levaduras *Saccharomyces cerevisiae* y *Schizosaccharomyces pombe*. La **FIGURA 5.5** resume las características más importantes. Los genomas de levadura de 12,5 Mb y 13,5 Mb tienen ~6.000 y ~5.000 genes, respectivamente. El marco de lectura abierto (ORF) es de ~1,4 kb; por lo tanto, ~70% del genoma está ocupado por regiones codificantes. La diferencia más importante entre ellos es que sólo el 5% de los genes de *S. cerevisiae* tiene intrones, en comparación con el 43% en *S. pombe*. La densidad de genes es alta; la organización en general es similar, aunque los espacios entre los genes son un poco más cortos en *S. cerevisiae*. Alrededor de la mitad de los genes identificados por secuenciación se conocían previamente o estaban relacionados con genes conocidos. Los restantes, eran previamente desconocidos, lo cual da algún indicio de la cantidad de nuevos tipos de genes que pueden descubrirse aún.

La identificación de marcos de lectura largos sobre la base de la secuencia es bastante precisa. Sin embargo, los ORF que codifican <100 aminoácidos no pueden identificarse solamente por la secuencia, debido a la alta aparición de falsos positivos. El análisis de la expresión génica sugiere que es probable que sean genes activos sólo ~300 de los 600 de tales ORF en *S. cerevisiae*.

Una vía poderosa para validar la estructura de genes es la de comparar la secuencia en especies estrechamente relacionadas: si un gen es activo, es probable que se conserve. Las comparaciones entre las secuencias de las cuatro especies relacionadas sugieren que 503 de los genes originalmente identificados en *S. cerevisiae* no tienen contrapartes en las otras especies y, por lo tanto, se deberían eliminar del catálogo. Esto reduce el número total de genes estimados para *S. cerevisiae* a 5.726.

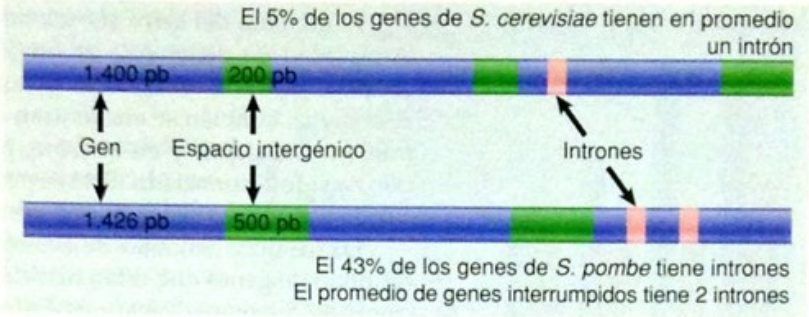


FIGURA 5.5 El genoma de *S. cerevisiae* de 13,5 Mb tiene 6.000 genes, casi todos sin interrupciones. El genoma de *S. pombe* de 12,5 Mb tiene 5.000 genes, casi la mitad tiene intrones. Los tamaños y el espaciamiento de los genes son bastante similares.

El genoma del DNA de *Caenorhabditis elegans* varía entre regiones ricas en genes y regiones en las cuales los genes se distribuyen de manera más dispersa. La secuencia total contiene ~18.500 genes. Sólo el ~42% de los genes tiene contrapartes putativas fuera de Nematoda.

El genoma de la mosca es más grande que el del gusano, pero existen menos genes en algunas especies (~14.000 en *D. melanogaster*) y más cantidad en otras (p. ej., ~23.000 en *D. persimilis*). El número de transcritos diferentes es, de alguna manera, mayor debido al proceso de corte y empalme alternativo. No se comprende por qué *C. elegans* –casi indiscutiblemente un organismo menos complejo– tiene un 30% más de genes que la mosca, pero puede ser debido a que *C. elegans* tiene un mayor promedio de genes, por familia de genes, de lo que tiene *D. melanogaster*; por lo tanto, el número de genes *únicos* de las dos especies es más similar. Una comparación de doce genomas de *Drosophila* revela que puede existir una variación bastante grande en la cantidad de genes entre especies estrechamente relacionadas. En algunos casos, existen varios miles de genes que son específicos de la especie. Esto enfatiza fuertemente la ausencia de una relación exacta entre el número de genes y la complejidad del organismo.

La planta *Arabidopsis thaliana* tiene un genoma intermedio entre el gusano y la mosca, pero tiene una cantidad de genes más grande (25.000) que cualquiera de los dos. Esto nuevamente demuestra la falta de una relación clara y también destaca la cualidad especial de las plantas, que pueden tener más genes (debido a duplicaciones ancestrales) que las células animales. Una mayoría del genoma de *Arabidopsis* se encuentra en segmentos duplicados, lo que sugiere que existió un ancestro con el genoma duplicado (un tetrahaploide). Sólo el 35% de los genes de *Arabidopsis* están presentes en copias simples.

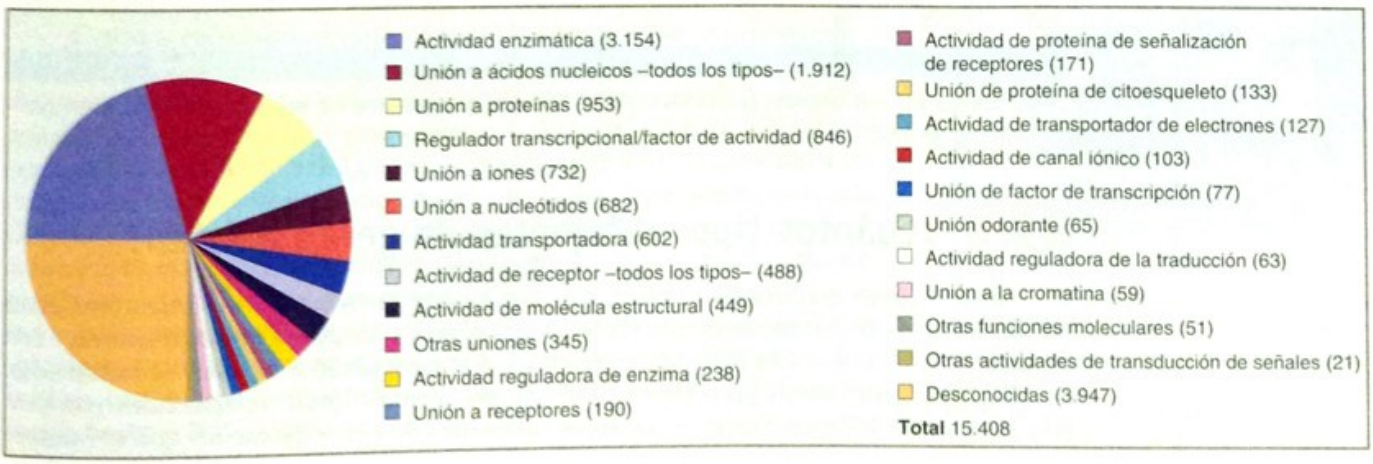


FIGURA 5.6 Funciones de los genes de *Drosophila* basados en la genómica comparativa de doce especies. Se desconocen las funciones de casi un cuarto de los genes de *Drosophila*. Adaptado de *Drosophila* 12 Genome Consortium, "Evolution of genes and genomes on the *Drosophila* phylogeny," *Nature* 450 (2007): 203–218.

El genoma del arroz (*Oryza sativa*) es ~4 veces más grande que el de *Arabidopsis*, pero la cantidad de genes sólo es un 50% mayor, probablemente ~40.000. El DNA repetitivo ocupa entre el 42 y 45% del genoma. Más del 80% de los genes que se encuentran en *Arabidopsis* también se encuentran en el arroz. De estos genes comunes, ~8.000 se encuentran en *Arabidopsis* y en el arroz, pero no en cualquier otro genoma bacteriano o animal que ha sido secuenciado. Éste es probablemente el conjunto de genes que codifican las funciones específicas de plantas, como la fotosíntesis.

De los doce genomas de *Drosophila* secuenciados, se puede formar una impresión de la cantidad de genes que están dedicados a cada tipo de función. La **FIGURA 5.6** separa las funciones en diferentes categorías. Entre los genes que se identificaron, se encuentran alrededor de 3.000 enzimas, ~900 factores de transcripción, y ~700 transportadores y canales iónicos. Alrededor de un cuarto de los genes codifican productos cuya función es aún desconocida.

El tamaño del polipéptido se incrementa desde los procariontes hasta los eucariontes. La *archaea M. jannaschi* y la bacteria *E. coli* tienen longitudes promedio de polipéptidos de 287 y 317 aminoácidos, respectivamente; mientras que *S. cerevisiae* y *C. elegans* tienen longitudes promedio de 484 y 442 aminoácidos, respectivamente. Los polipéptidos grandes (>500 aminoácidos) son poco frecuentes en las bacterias, pero abarcan un componente significativo (~1/3) en los eucariontes. El aumento en la longitud se debe a la adición de dominios extras, con cada dominio típico constituido por 100 a 300 aminoácidos. Sin embargo, el incremento en el tamaño del polipéptido, es responsable de sólo una parte pequeña del incremento en el tamaño del genoma.

Otro conocimiento profundo de la cantidad de genes se obtiene mediante el recuento del número de genes que codifican proteínas que se expresan. Sobre la base de la estimación del número de diferentes especies de mRNA que pueden contarse en una célula, se podría concluir que en los vertebrados, la célula promedio expresa de entre ~10.000 a ~20.000 genes. La existencia de solapamientos significativos entre las poblaciones de mRNA en diferentes tipos de célula sugeriría que la cantidad total de genes expresados para el organismo debería estar dentro del mismo orden de magnitud. La estimación de la cantidad de genes humanos totales de 20.000 a 25.000 (véase la *Sección 5.5, El genoma humano tiene menos genes de lo esperado originalmente*) podría implicar que una proporción significativa del número total de genes en realidad se expresa en cualquier célula.

- **mRNA monocistrónico:** mRNA que codifica un solo polipéptido.
- **mRNA policistrónico:** mRNA que incluye regiones codificantes que representan a más de un gen.

Los genes eucariontes se transcriben individualmente: cada gen produce un mensajero **monocistrónico**. Existe una excepción general de esta regla: en el genoma de *C. elegans*, el 15% de los genes están organizados en unidades **policistrónicas** (que están asociadas con el uso del corte y empalme *trans* para permitir la expresión de los genes hacia el extremo 3′ en estas unidades; véase la *Sección 28.12, Las reacciones de corte y empalme en la región* trans *usan secuencias cortas de RNA*).

CONCEPTOS CLAVE

- Existen 6.000 genes en la levadura; 18.500 en un gusano; 13.600 en una mosca; 25.000 en la planta pequeña *Arabidopsis*; y probablemente de 20.000 a 25.000 en los seres humanos y los ratones.

REVISIÓN DE CONCEPTOS

¿Por qué en los eucariontes multicelulares el tamaño del genoma no es un buen indicador del número de genes?

5.4 ¿Cuántos tipos diferentes de genes existen?

Algunos genes son únicos; otros, pertenecen a familias en las que sus miembros están relacionados –pero no suelen ser idénticos–. La proporción de genes únicos disminuye con el tamaño del genoma y la proporción de genes que pertenecen a familias se incrementa.

Algunos genes están presentes en más de una copia o están relacionados con otro gen, de manera tal que el número de tipos diferentes de genes es menor que el número total de genes. Se puede dividir al número total de genes en conjuntos que tengan miembros relacionados, definidos a través de la comparación de sus exones. (Una familia de genes surge por duplicación de un gen ancestral seguido por una acumulación de cambios en la secuencia entre las copias. Muy frecuentemente los miembros de una familia

están relacionados entre sí, pero no son idénticos.) El número de tipos de genes se calcula al sumar la cantidad de genes únicos (para los cuales no existe otro gen relacionado) a la cantidad de familias que tienen dos o más miembros.

La **FIGURA 5.7** compara el número total de genes con el número de familias distintivas en cada uno de los seis genomas. En las bacterias, la mayoría de los genes son únicos; por lo tanto, el número de familias distintivas es cercano al número total de genes. La situación es diferente incluso en el eucarionte unicelular *S. cerevisiae*, para el cual existe una proporción significativa de genes repetidos. El efecto más notable es que el número de genes se incrementa abruptamente en los eucariontes superiores, pero la cantidad de familias génicas no cambia demasiado.

La **FIGURA 5.8** muestra que la proporción de genes únicos disminuye abruptamente según el tamaño del genoma. Cuando los genes están agrupados en familias, el número de sus miembros en una familia es pequeño en las bacterias y en los eucariontes unicelulares, pero grande en los eucariontes multicelulares. Gran parte del tamaño extra del genoma de *Arabidopsis* se puede explicar por la presencia de familias con más de cuatro miembros.

Si cada gen se expresa, el número total de genes representará el número total de polipéptidos necesarios para constituir un organismo (proteoma). Sin embargo, existen dos condiciones que causan que el proteoma sea diferente del número total de genes. Primero, los genes pueden ser duplicados y, en consecuencia, algunos de ellos codifican el mismo polipéptido (aunque pueden expresarse en tiempos y lugares diferentes) y otros pueden codificar polipéptidos relacionados que nuevamente cumplen la misma función en tiempos y lugares diferentes. Segundo, el proteoma puede ser más grande que el número de genes, debido a que los genes pueden producir más de un polipéptido mediante el proceso de corte y empalme alternativo.

¿Qué es el proteoma central –el número básico de tipos diferentes de polipéptidos en el organismo? Aunque es dificultoso de estimar, debido a la posibilidad de corte y empalme alternativo, una estimación mínima estaría dada por el número de familias génicas, las cuales varían desde 1.400 en la bacteria, hasta ~4.000 en la levadura, y desde 11.000 hasta 14.000 para la mosca y el gusano.

¿Cuál es la distribución del proteoma según el tipo de proteína? Las 6.000 proteínas del proteoma de levadura incluyen 5.000 proteínas solubles y 1.000 proteínas de membrana. Alrededor de la mitad de las proteínas son citoplasmáticas, un cuarto se encuentran en el núcleo, y el resto están repartidas entre la mitocondria y el retículo endoplasmático (RE)/sistema de Golgi.

¿Cuántos genes son comunes a todos los organismos (o grupos como las bacterias o los eucariontes multicelulares), y cuántos son específicos del taxón? La **FIGURA 5.9** muestra la comparación entre los genes de la mosca, los del gusano (otro eucarionte multicelular) y los de la levadura (un eucarionte unicelular). Los genes que codifican los polipéptidos correspondientes en diferentes organismos se denominan **ortólogos** (véase la *Sección 3.9 Los miembros de una familia de genes tienen una organización común*). Funcionalmente, se suele considerar que dos genes en diferentes organismos son ortólogos si sus secuencias son similares en más del 80% de la longitud. Según este criterio, ~20% de los genes de la mosca presentan ortólogos, tanto en la levadura como en el gusano. Estos genes son, probablemente, requeridos por todos los eucariontes. La proporción se incrementa hasta el 30% cuando se compara los genes de la mosca con los del gusano, probablemente al representar la adición de funciones de genes que son comunes a los eucariontes multicelulares. Esto incluso deja una proporción importante de genes que codifican proteínas, que son requeridas específicamente por las moscas o los gusanos, respectivamente.

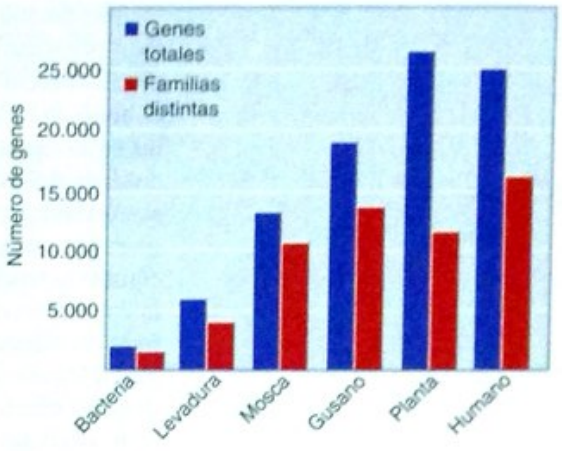


FIGURA 5.7 Muchos genes están duplicados y, como resultado, el número de familias génicas diferentes es mucho menor que el número total de genes. El histograma compara el número total de genes con el número de familias génicas distintas.

	Genes únicos	Familias con 2-4 miembros	Familias con >4 miembros
H. influenzae	89%	10%	1%
S. cerevisiae	72%	19%	9%
D. melanogaster	72%	14%	14%
C. elegans	55%	20%	26%
A. thaliana	35%	24%	41%

FIGURA 5.8 La proporción de genes que están presentes en copias múltiples se incrementa con el tamaño del genoma en los eucariontes multicelulares.

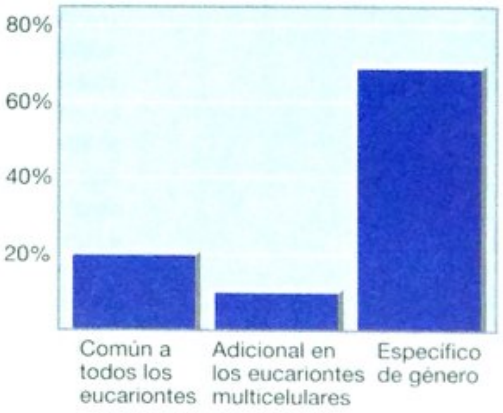


FIGURA 5.9 El genoma de la mosca se puede dividir en los genes que se encuentran, probablemente, en todos los eucariontes; los genes adicionales que se encuentran, probablemente, en todos los organismos multicelulares; y los genes, que son más específicos a subgrupos de especies que incluyen a las moscas.

genes ortólogos (ortologos): Genes que están relacionados en diferentes especies.

Se puede deducir una estimación mínima del tamaño de un proteoma perteneciente a un organismo a partir de la cantidad y de las estructuras de los genes, y también se puede medir directamente el tamaño del proteoma celular o perteneciente al organismo mediante el análisis del contenido total de polipéptidos de la célula o del organismo. Mediante tales estrategias, se identificaron algunas proteínas que no se sospechaban sobre la base del análisis del genoma; esto condujo a la identificación de genes nuevos. Se utilizaron diversos métodos para el análisis a gran escala de proteínas. La espectrometría de masa se puede usar para separar e identificar proteínas en una mezcla obtenida directamente a partir de células o tejidos. Las proteínas híbridas que llevan etiquetas pueden obtenerse mediante la expresión de los cDNA sintetizados al ligar las secuencias de los ORF a vectores de expresión adecuados que incorporen las secuencias por afinidad de etiquetas. Esto permite que a través del análisis de matriz se analicen los productos. Estos métodos también pueden ser efectivos al comparar las proteínas de dos tejidos –por ejemplo, un tejido de un individuo sano con otro de un paciente con enfermedad–, para precisar las diferencias.

Además de los genes funcionales, también existen copias de genes que se convirtieron en no funcionales (identificados como tales mediante interrupciones en sus secuencias que codifican proteínas). Éstos se denominan *pseudogenes* (véase la *Sección 6.1, Introducción*). La cantidad de pseudogenes puede ser grande. En los genomas humano y de ratón, la cantidad de pseudogenes es ~10% de la cantidad de genes activos (o potencialmente activos) (véase la *Sección 4.8, La conservación de la organización del genoma contribuye a identificar genes*).

CONCEPTOS CLAVE

- La suma del número de genes únicos y del número de familias de genes es una estimación del número de tipos de genes.
- Se puede estimar el tamaño mínimo del proteoma a partir de la cantidad de tipos de genes.

REVISIÓN DE CONCEPTOS

¿Por qué el tamaño del proteoma no es una estimación precisa del número de tipos de genes?

5.5 El genoma humano tiene menos genes de lo esperado originalmente

El genoma humano fue el primer genoma de vertebrado en ser secuenciado. Esta tarea colosal reveló una riqueza de información acerca de la composición genética de nuestra especie y de la evolución de los genomas en general. El entendimiento se profundizó aún más por la capacidad de comparar la secuencia del genoma humano con otras secuencias de genomas de vertebrados.

Los genomas de mamíferos suelen encontrarse en un intervalo de tamaño acotado ~3 × 10^9 pb (véase la *Sección 4.5, ¿Por qué algunos genomas son tan grandes?*). El genoma de ratón es ~14% más pequeño que el genoma humano, probablemente debido a que ha tenido una tasa de deleción más alta. Los genomas contienen genes y familias de genes similares: la mayoría de los genes tiene un ortólogo en otro genoma, pero con diferencias en el número de miembros de una familia, especialmente en aquellos casos para los cuales las funciones son específicas de la especie (véase la *Sección 4.8, La conservación de la organización del genoma contribuye a identificar genes*). Originalmente, se estimó que tenía ~30.000 genes; en la actualidad, se cree que el genoma de ratón tiene el mismo número de genes que el genoma humano, de 20.000 a 25.000. Los 23.000 genes que codifican proteínas están acompañados por ~3.000 genes que representan los RNA que no codifican proteínas; en general, éstos son pequeños (aparte de los RNA ribosómicos). Casi la mitad de estos genes codifica RNA de transferencia. Además de los genes activos, se han identificado ~1.200 pseudogenes.

El genoma humano (haploide) contiene 22 autosomas, además de los cromosomas X e Y. El tamaño de los cromosomas varía desde 45 hasta 279 Mb de DNA, siendo el tamaño total del genoma de 3.286 Mb (~3,3 × 10^9 pb). Sobre la base de la estructura del cromosoma, el genoma se puede dividir en regiones de eucromatina (que contiene muchos genes activos) y heterocromatina, con una densidad de genes activos mucho menor (véase la *Sección 23.5, La cromatina se divide en eucromatina y heterocromatina*). La eucromatina abarca la mi-

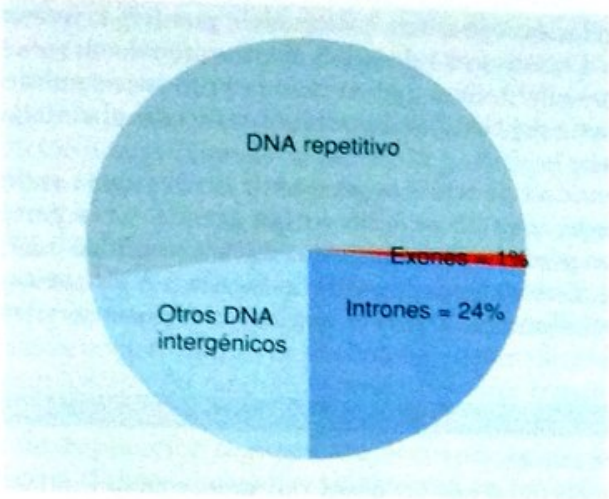


FIGURA 5.10 Los genes ocupan el 25% del genoma humano, pero las secuencias que codifican proteínas representan sólo una parte pequeña de esta fracción.

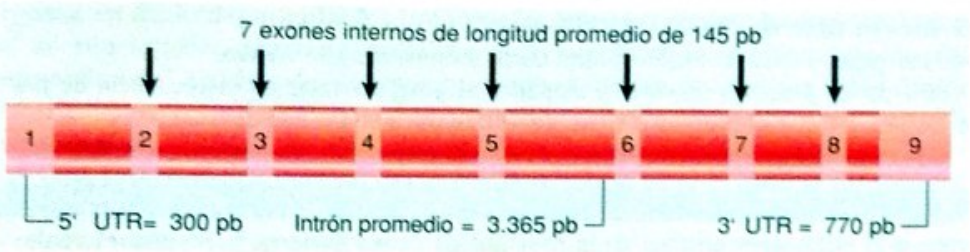


FIGURA 5.11 El gen humano promedio tiene 27 kb de longitud y presenta nueve exones; en general, abarca dos exones largos en cada extremo y siete exones internos. Los UTR en los exones terminales son las regiones no traducidas (no codificantes) en cada extremo del gen. (Esto está basado en el promedio. Algunos genes son extremadamente largos, lo que da como resultado una longitud media de 14 kb con siete exones.)

yor parte del genoma, ~2,9 × 10^9 pb. La secuencia del genoma que ha sido identificada representa ~90% de la eucromatina. Además de proporcionar información sobre el contenido del genoma, la secuencia también identifica características que pueden ser de importancia estructural (véase la *Sección 23.6, Los cromosomas tienen patrones de bandeo*).

La **FIGURA 5.10** muestra que una proporción muy pequeña (~1%) del genoma humano está constituido por los exones que realmente codifican polipéptidos. Los intrones que constituyen las secuencias remanentes de genes que codifican proteínas aportan el total de DNA relacionado con la producción de proteínas, y alcanza ~25%. Como se muestra en la **FIGURA 5.11**, el gen humano promedio es de 27 kb de largo, con nueve exones que incluyen una secuencia codificante total de 1.340 pb. La secuencia codificante promedio es, por lo tanto, sólo el 5% de la longitud de un gen promedio que codifica una proteína.

Mediante cualquier medición, la cantidad total de genes humanos –de 20.000 a 25.000– es mucho menos de lo que se esperaba originalmente, ya que la mayoría de las estimaciones, antes de que se secuenciara el genoma, eran de ~100.000. Esto demuestra un incremento relativamente pequeño por encima de las moscas y de los gusanos (~13.000 y ~18.500, respectivamente), sin mencionar al de la planta *Arabidopsis thaliana* (~25.000) (véase la Figura 5.2). Sin embargo, no debería ser particularmente sorprendente el hecho de que no se requiera un gran número de genes adicionales para formar un organismo más complejo. La diferencia en las secuencias de DNA entre los genomas humano y de chimpancé es extremadamente pequeña: existe > 99% de similitud; por lo tanto, es claro que las funciones y las interacciones entre un conjunto similar de genes puede producir resultados muy diferentes. Las funciones de los grupos específicos de genes pueden ser especialmente importantes, debido a que las comparaciones detalladas de los genes ortólogos en los seres humanos y en los chimpancés sugieren que ha existido una evolución rápida de cierta clases de genes, incluyendo algunas involucradas en el desarrollo temprano, el olfato y la audición –todas funciones que son relativamente especializadas en estas especies.

La cantidad de genes que codifican proteínas es menor que la cantidad de polipeptidos potenciales, debido a los mecanismos de corte y empalme alternativo, a la selección de promotor alternativo y a la selección del sitio de poli (A) alternativo que pueden producir diversos polipéptidos a partir de un mismo gen (véase la *Sección 28.11, El corte y empalme alternativo involucran el uso diferencial de las uniones de corte y empalme*). El alcance del corte y empalme alternativo es mayor en los seres humanos que en las moscas y en los gusanos; este mecanismo puede afectar hasta el 60% de los genes; por lo tanto, el incremento

en el tamaño del proteoma humano, en relación con otros eucariontes, puede ser mayor que el incremento en el número de genes. Una muestra de genes de dos cromosomas sugiere que la proporción de corte y empalme alternativo, que en realidad produce cambios en la secuencia polipeptídica, puede ser hasta del 80%. Esto podría incrementar el tamaño del proteoma de 50.000 a 60.000 miembros.

En términos de la diversidad de la cantidad de familias génicas, la discrepancia entre los seres humanos y otros eucariontes puede, sin embargo, no ser tan grande. Gran parte de los genes humanos pertenecen a familias génicas. Un análisis de ~25.000 identificó 3.500 genes únicos y 10.300 genes emparentados. Como se puede ver en la Figura 5.7, esto se extrapola a un número de familias génicas sólo levemente mayor que la de los gusanos y las moscas.

CONCEPTOS CLAVE

- Sólo el 1% del genoma humano consiste en exones.
- Los exones abarcan ~5% de cada gen, de manera tal que los genes (los exones junto con los intrones) comprenden el ~25% del genoma.
- El genoma humano tiene de 20.000 a 25.000 genes.
- El ~60% de los genes humanos experimentan corte y empalme alternativo.
- Hasta el ~80% de los procesos de corte y empalme alternativo cambian la secuencia de proteínas, de manera tal que el proteoma tiene de ~50.000 a 60.000 miembros.

REVISIÓN DE CONCEPTOS

¿Cómo se obtuvo la estimación original de la cantidad de genes humanos? ¿Por qué el verdadero número es mucho menor?

5.6 ¿Cómo se distribuyen los genes y otras secuencias en el genoma?

¿Se distribuyen en forma uniforme los genes en el genoma? Algunos cromosomas son relativamente pobres en genes y tienen > 25% de sus secuencias como "desiertos" –regiones más largas que 500 kb, donde no existen los ORF. Incluso, la mayoría de los cromosomas ricos en genes tiene > 10% de sus secuencias como desiertos. Por lo tanto, de manera global, ~20% del genoma humano consiste de desiertos que no tienen genes que codifican proteínas.

Las secuencias repetitivas representan ~50% del genoma humano, como se observa en la **FIGURA 5.12**. Las secuencias repetitivas pertenecen a cinco clases:

- Los transposones (ya sean activos o inactivos) representan a la gran mayoría (45% del genoma). Todos los transposones se encuentran en copias múltiples.
- Los pseudogenes procesados (~3.000 en total, representan ~0,1% del DNA total). (Éstas son secuencias que surgen por inserción de una copia de DNA transcripta de manera inversa a partir de una secuencia de mRNA en el genoma; véase la *Sección 6.1, Introducción*.)
- Las repeticiones de secuencia simples (DNA altamente repetitivo como $(CA)_n$, representan ~3%).
- Las duplicaciones segmentarias (segmentos de 10 a 300 kb que han sido duplicados en una región nueva) representan ~5%. Sólo una minoría de estas duplicaciones se encuentran en el mismo cromosoma; el resto de las duplicaciones, están en cromosomas diferentes.
- Las repeticiones en tándem forman bloques de un tipo de secuencia (que se encuentran, en especial, en los centrómeros y los telómeros).

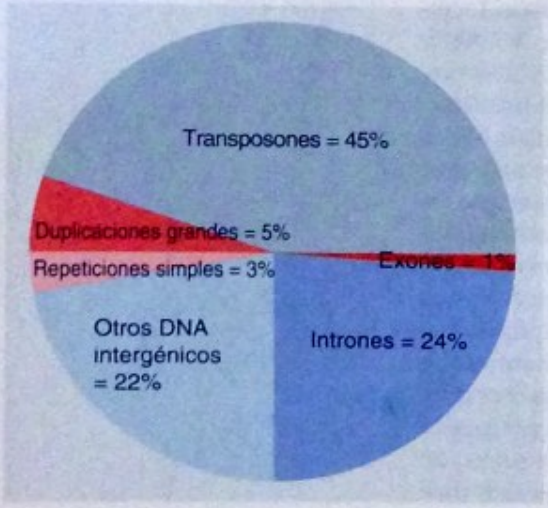


FIGURA 5.12 El componente más grande del genoma humano consiste en los transposones. Otras secuencias repetitivas incluyen duplicaciones grandes y repeticiones simples.

La secuencia del genoma humano enfatiza la importancia de los transposones. (Muchos transposones tienen la capacidad de replicarse a sí mismos y de insertarse en ubicaciones nuevas.) Pueden funcionar exclusivamente como elementos de DNA o pueden tener una forma activa, que es RNA (véase el *Capítulo 21, Transposones, retrovirus y retrotransposones*). En la Figura 21.35, se resume su distribución en el genoma humano. La mayoría de los transposo-

nes en el genoma humano son no funcionales; en realidad, muy pocos son activos. Sin embargo, la alta proporción del genoma ocupado por estos elementos indica que pueden cumplir un papel activo en la configuración del genoma. Una característica interesante es que algunos genes presentes se originaron como transposones y evolucionaron hasta su condición actual después de perder la capacidad para transponerse. Casi 50 genes parecen haberse originado de esta manera.

La duplicación segmentaria en su forma más simple involucra la duplicación en tándem de algunas regiones dentro de un cromosoma (en general, debido a un acontecimiento de recombinación aberrante en la meiosis; véase la *Sección 6.6, El entrecruzamiento desigual reordena los grupos de genes*). En muchos casos, sin embargo, las regiones duplicadas se encuentran en cromosomas diferentes, lo que implica que existió originalmente una duplicación en tándem seguida por una translocación de una copia a un sitio nuevo, o que la duplicación surgió por algún mecanismo del todo diferente. El caso extremo de una duplicación segmentaria es cuando se duplica un genoma entero, en cuyo caso el genoma diploide inicial se transforma en tetraploide. A medida que las copias duplicadas desarrollan diferencias una de otra, el genoma puede convertirse nuevamente y de forma gradual en diploide, aunque las homologías entre las copias divergentes dejen evidencia del acontecimiento. Esto es especialmente frecuente en los genomas de las plantas. El estado actual del análisis del genoma humano identifica muchas regiones duplicadas individuales, pero no indica si existió una duplicación de genoma completo en el linaje de los vertebrados.

Una característica curiosa del genoma humano es la presencia de secuencias que no parecen tener funciones codificantes pero que, no obstante, muestran una conservación evolutiva mayor que el nivel basal. Como se detectó por comparación con otros genomas (como el genoma de ratón), éstas representan alrededor del 5% del genoma total. Estas secuencias ¿están asociadas de alguna manera funcional con las secuencias que codifican proteínas? Su densidad en el cromosoma 18 es la misma que en cualquier lugar en el genoma, aunque el cromosoma 18 tiene una concentración significativamente menor de genes que codifican proteínas. Esto sugiere indirectamente que su función no está conectada con la estructura o con la expresión de los genes que codifican proteínas.

CONCEPTOS CLAVE

- Las secuencias repetidas (presentes en más de una copia) representan >50% del genoma humano.
- La gran masa de secuencias repetidas consiste en copias de transposones no funcionales.
- Existen muchas duplicaciones de regiones cromosómicas grandes.

REVISIÓN DE CONCEPTOS

¿Qué mecanismos pueden producir duplicaciones en tándem de segmentos cromosómicos? ¿Qué mecanismo puede explicar la duplicación desplazada que se encuentra en un cromosoma diferente?

5.7 El cromosoma Y tiene diversos genes específicos del sexo masculino

La secuencia del genoma humano ha extendido de manera significativa la comprensión de la función de los cromosomas sexuales. En general, se cree que los cromosomas X e Y descendieron de un par de autosomas (muy ancestrales) común. Su evolución ha involucrado un proceso en el cual el cromosoma X retuvo la mayoría de los genes originales, mientras que el cromosoma Y perdió la mayoría de ellos.

El cromosoma X es similar a los autosomas a tal grado que las mujeres tienen dos copias, y se puede producir la recombinación entre éstas. La densidad de genes en el cromosoma X es comparable con la densidad de genes en otros cromosomas.

El cromosoma Y es mucho más pequeño y tiene mucho menos genes que el cromosoma X. Su papel único resulta del hecho de que sólo los varones tienen el cromosoma Y, de los cuales existe una sola copia; por lo tanto, los loci son efectivamente haploides, en lugar de diploide como todos los otros genes humanos.

MÉTODOS Y TÉCNICAS

Rastreo de la historia humana a través del cromosoma Y

Debido a que el cromosoma Y no sufre recombinación a lo largo de la mayoría de su longitud, los marcadores genéticos en el cromosoma Y están completamente ligados y permanecen juntos a medida que el cromosoma se transmite de generación en generación. Por lo tanto, se puede rastrear la relación genética entre los cromosomas Y, debido a que los cromosomas que están más estrechamente relacionados compartirán más alelos a lo largo de su longitud de lo que lo harán los cromosomas relacionados de forma más distante. El conjunto de alelos de un cromosoma particular se denomina **haplotipo**. Para muchos estudios genealógicos del cromosoma Y, los polimorfismos de secuencias repetidas simples (SSR) son convenientes debido a su tasa de mutación relativamente alta, producto de los errores de replicación, y al gran número de alelos. La lógica es que los cromosomas Y, con haplotipos que comparten alelos en cada uno de los 20–30 SSR a través del cromosoma, deben descender del mismo cromosoma Y ancestral en el pasado muy reciente. Para haplotipos que difieren en un locus único, la relación genética es menos cercana; para aquellos que difieren en dos loci, es aún menos cercana, y así sucesivamente. Esta lógica simple es la base para el rastreo de la historia de la población a través de los polimorfismos del cromosoma Y. Los haplotipos que comparten muchos alelos tienen un cromosoma Y ancestral común más reciente que los haplotipos que comparten menos alelos. Además, debido a que se puede estimar la tasa de mutación de los SSR, se puede deducir el tiempo en el que existió el cromosoma ancestral. Este razonamiento forma las bases para las estimaciones de que el ancestro común más reciente de todos los cromosomas Y humanos existió hace 50.000 años. Tales estimaciones no son altamente precisas, y se deben formular muchas presunciones. Otros estudios que utilizan diferentes marcadores dieron una estimación de 150.000 años. Las razones para la discrepancia son aún inciertas. No obstante, se puede aprender mucho acerca de la historia de la población humana mediante el estudio del cromosoma Y. Por ejemplo, el legado de Genghis Khan se puede investigar mediante el rastreo del linaje del cromosoma Y.

En su máximo alcance, desde China hasta Rusia a través del Oriente Medio y luego dentro de Europa del Este, el Imperio Mongol del siglo XIII abarcó el Imperio de tierras más grande que la historia ha conocido. Su fundador nació con el nombre de Temujin, alrededor de año 1162. Siendo un hombre joven organizó una confederación de tribus, quienes alrededor del año 1200 tomaron sus pequeños ponis mongoles equipados con sus altas sillas de montar de madera y estribos y, armados con arco y flecha, comenzaron a conquistar a sus vecinos. Después, Temujin adoptó el nombre de Genghis Khan, que significa *gobernante universal*. A menudo fue despiadado: exterminó a los hombres y niños de las ciudades rebeldes y secuestró a las mujeres y niñas. En respuesta a una pregunta acerca de la fuente de felicidad, se hizo célebre por haber dicho: "La mayor felicidad es la de vencer a tus enemigos, la de cazarlos antes que tú, la de robarles sus riquezas, la de ver a sus seres queridos bañados en lágrimas, la de asir los senos de sus viudas e hijas". A través de sus múltiples esposas, concubinas, y las innumerables conquistas sexuales no reconocidas, Genghis Khan y sus descendientes fueron muy prolíficos. Su hijo mayor, Tushi, tuvo 40 hijos reconocidos, y su nieto Kubilai Khan (bajo quien el Imperio Mongol alcanzó su máxima extensión) tuvo 22 hijos reconocidos.

A pesar de que el legado de Genghis Khan está bien registrado en la historia, apenas se esperaba que se pudiera demostrar mediante los estudios del cromosoma Y. Pero, los estudios del genotipo de 32 marcadores a lo largo del cromosoma Y, de 2.123 muestras de hombres provenientes de una gran región de Asia, produjeron el resultado notable que se observa en la FIGURA B5.1. Cada círculo representa una muestra de la población, con su área proporcional al tamaño de la muestra. Los sectores rojos indican la frecuencia relativa de un grupo de haplotipos casi idénticos del cromosoma Y, mientras que los sectores blancos representan la frecuencia relativa de otros haplotipos que son genéticamente mucho más diversos. El ancestro común más reciente de los haplotipos más estrechamente relacionados se estima que existió hace 1.000 ± 300 años. Además, la región geográfica

Haplotipo: Combinación particular de alelos en una región definida de un cromosoma de la que resulta un genotipo en miniatura. Anteriormente se lo utilizaba para describir las combinaciones de alelos del complejo mayor de histocompatibilidad (CMH); actualmente puede usarse para describir combinaciones determinadas de RFLP, SNP u otros marcadores.

Durante muchos años, se pensó que el cromosoma Y casi no portaba genes, excepto de uno a unos pocos que determinan la masculinidad. La gran mayoría del cromosoma Y (>95% de su secuencia) no sufre entrecruzamiento con el cromosoma X, lo que conduce a la visión de que podría no contener genes activos debido a que no habría manera de prevenir la acumulación de mutaciones deletéreas. Esta región está flanqueada por regiones pseudoautosómicas cortas que se intercambian frecuentemente con el cromosoma X durante la meiosis masculina. Originalmente se la denominó región *no recombinante*, pero en la actualidad se la denomina *región específica masculina*.

La secuenciación detallada del cromosoma Y muestra que la región específica masculina contiene tres tipos de secuencias, como se ilustra en la FIGURA 5.13.

- Las *secuencias X transpuestas* consisten en un total de 3,4 Mb, que comprenden algunos bloques grandes que son el producto de una transposición de la banda q21 en el cromosoma X, hace aproximadamente 3 o 4 millones de años atrás. Esto es

en la cual se agrupan los haplotipos más estrechamente relacionados, está incluida en gran parte dentro del Imperio Mongol (sombreado). La única excepción es la población 10, compuesta por la etnia Hazara de Pakistán. Esto proporciona una pista para el origen de los cromosomas Y estrechamente relacionados, debido a que Los Hazara se consideran a sí mismos de origen Mongol, y muchos aclaman ser descendientes directos de la línea masculina de Genghis Khan. Cualquiera sea su origen, los cromosomas Y estrechamente relacionados se encuentran en alrededor del 8% de los hombres a través de una gran región de Asia (poblaciones 1–16). En principio, la prueba directa de la conexión con Genghis Khan podría obtenerse por determinación del haplotipo del cromosoma Y en el material recuperado de su tumba. Murió en 1227 a partir de las lesiones producto de una caída del caballo, y se dice que ha sido enterrado cerca de su lugar de nacimiento.

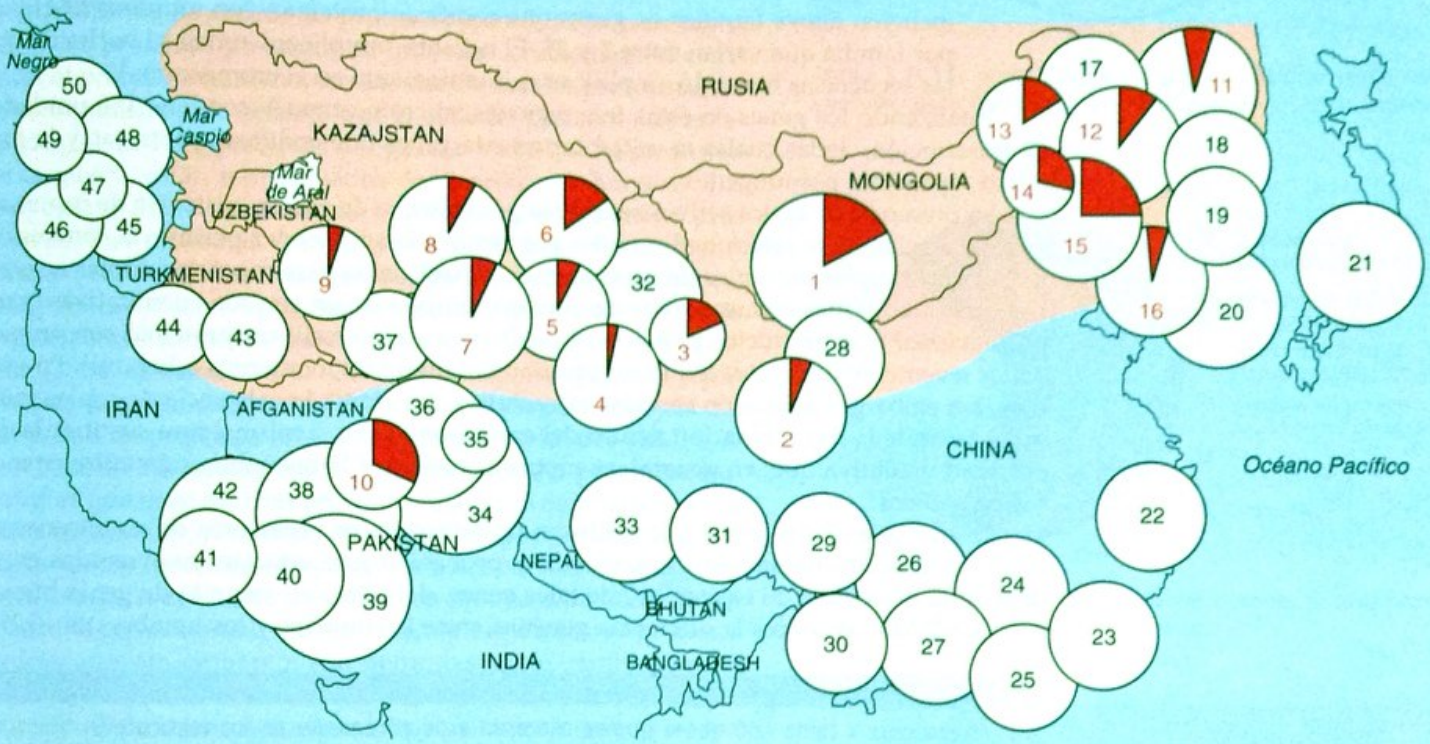


FIGURA B5.1 Distribución de los haplotipos del cromosoma Y (rojo), que se presume descienden de Genghis Khan o sus parientes masculinos cercanos, entre las poblaciones cercanas y en la frontera del Imperio Mongol antiguo. Los grupos de poblaciones específicas son: (1) mongol, (2) han [Gansu], (3) kazak chino, (4) han [Xinkiang], (5) xibe, (6) uigur, (7) kirguis, (8) kazak, (9) uzbekeo, (10) hazara, (11) hezhen, (12) daur, (13) evenki, (14) han [mongol del interior], (15) mongol del interior, (16) manchú, (17) oroqen, (18) han [Heilongjiang], (19) coreano chino, (20) coreano, (21) japonés, (22) she, (23) han [Guangdong], (24) yao [Liannan], (25) li, (26) buyí, (27) yao [Bama], (28) hui, (29) han [Sichuan], (30) hani, (31) qiang, (32) uigur chino, (33) tibetano, (34) burusho (35) balti, (36) kalasha, (37) tayikos, (38) balochis, (39) parsi, (40) makrani negroide, (41) makrani balochis, (42) brahui, (43) turcomano, (44) kurdo, (45) azeni, (46) armenio, (47) lezguiano, (48) georgiano, (49) osetio, (50) esvano. [Adaptado de Zerjal, T. *Am. J. Hum. Genet.* 72 (2003): 717–721.]

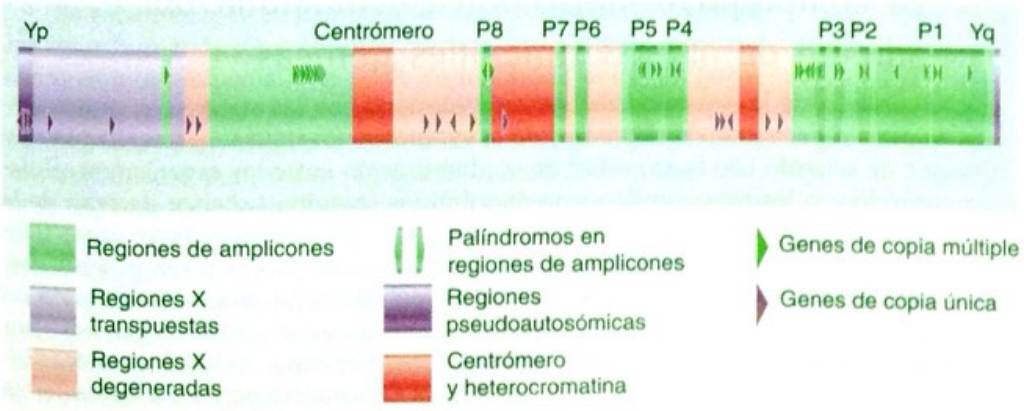


FIGURA 5.13 El cromosoma Y consiste en regiones X transpuestas, regiones X degeneradas y amplicones. Las regiones X transpuestas y X degeneradas tienen dos y catorce genes de copia única, respectivamente. Los amplicones tienen ocho palíndromos grandes (P1–P8), que contienen nueve familia de genes. Cada familia contiene, al menos, dos copias.

específico del linaje humano. Estas secuencias no se recombinan con el cromosoma X, y gran parte de ellas se han tornado inactivas. En la actualidad contienen sólo dos genes activos.

- Los *segmentos X degenerados* del cromosoma Y son secuencias que tienen un origen en común con el cromosoma X (retrocediendo hasta el autosoma común a partir del cual descendieron tanto el cromosoma X como el Y) y contienen genes o pseudogenes relacionados con los genes ligados al X. Existen 14 genes activos y 13 pseudogenes. Los genes activos, en cierto modo, hasta ahora han desafiado la tendencia de los genes a ser eliminados de las regiones cromosómicas que no pueden recombinarse por meiosis.
- Los *segmentos de amplicones* tienen una longitud total de 10,2 Mb y se repiten internamente en el cromosoma Y. Existen ocho bloques palindrómicos grandes. Ellos incluyen nueve familias de genes que codifican proteínas, con números de copias por familia que varían entre 2 y 35. El nombre "amplicón" refleja el hecho de que las secuencias han sido amplificadas internamente en el cromosoma Y.

Totalizando los genes en estas tres regiones, el cromosoma Y contiene 156 unidades de trascripción, de las cuales la mitad representa genes que codifican proteínas y la otra mitad representa pseudogenes.

La presencia de genes activos se explica por el hecho de que la existencia de copias de genes estrechamente relacionados en los segmentos de amplicones permite la conversión génica entre copias múltiples de un gen, para ser utilizadas para regenerar copias activas. Las necesidades más comunes para las copias múltiples de un gen son cuantitativas (para proporcionar más productos proteicos) o cualitativas (para codificar proteínas con propiedades levemente diferentes o que se expresan en diferentes momentos o lugares). En este caso, sin embargo, la función esencial es evolutiva. En efecto, la existencia de copias múltiples permite la recombinación dentro del cromosoma Y en sí mismo para sustituir la diversidad evolutiva que, en general, es proporcionada por la recombinación entre cromosomas alélicos.

La mayoría de los genes que codifican proteínas en los segmentos de amplicones se expresan específicamente en los testículos y probablemente estarían involucradas en el desarrollo masculino. Si existen ~60 de tales genes, del total del conjunto de genes humanos de ~25.000, entonces la diferencia genética entre las mujeres y los hombres es ~0,2%.

CONCEPTOS CLAVE

- El cromosoma Y tiene ~60 genes que se expresan específicamente en los testículos.
- Los genes específicos del sexo masculino están presentes en copias múltiples en segmentos cromosómicos repetidos.
- La conversión génica entre las copias múltiples permite que los genes activos se mantengan durante la evolución.

REVISIÓN DE CONCEPTOS

¿Por qué se asumió en un principio que el cromosoma Y humano podría tener muy pocos genes activos?

5.8 La complejidad morfológica evoluciona por adición de nuevas funciones génicas

La comparación de la secuencia del genoma humano con las secuencias que se encuentran en otras especies revela el proceso de la evolución. La **FIGURA 5.14** analiza los genes humanos de acuerdo con la amplitud de su distribución entre los organismos celulares. Comenzando con los más ampliamente distribuidos (esquina superior derecha de la figura), el 21% de los genes son comunes a los eucariontes y procariontes. Estos genes tienden a codificar proteínas que son esenciales a todas las formas vivientes –en general, metabolismo básico, replicación, transcripción y traducción. Moviéndose en el sentido de las agujas del reloj, otro 32% de los genes se encuentran en los eucariontes en general –por ejemplo, se pueden encontrar en la levadura. Éstos genes tienden a codificar proteínas involucradas en funciones que son comunes a las células eucariontes pero no a las bacterias

–por ejemplo, pueden pertenecer a orgánulos específicos o componentes del citoesqueleto. Otro 24% de los genes se encuentran generalmente en los animales. Éstos incluyen los genes necesarios para la multicelularidad y el desarrollo de diferentes tipos de tejidos. El 22% de los genes son únicos de vertebrados. Éstos mayormente codifican proteínas de los sistemas inmunológico y nervioso; ellos codifican muy pocas enzimas, consistente con la idea de que las enzimas tienen orígenes ancestrales, y que las vías metabólicas se originaron temprano en la evolución. Por lo tanto, se puede ver que la evolución de morfologías y especializaciones más complejas requiere la adición de grupos de genes que representan las funciones nuevas necesarias.

Una forma de definir las proteínas esenciales es la de identificar las proteínas presentes en todos los proteomas. Al comparar el proteoma humano, en más detalle con los proteomas de otros organismos, se observa que el 46% del proteoma de la levadura, el 43% del proteoma del gusano y el 61% del proteoma de la mosca, están representados en el proteoma humano. Un grupo clave de ~1.300 proteínas está presente en los cuatro proteomas. Las proteínas comunes son proteínas "de mantenimiento" fundamentales requeridas para funciones esenciales, que pertenecen a los tipos que se resume en la FIGURA 5.15. Las principales funciones se refieren a la transcripción y la traducción (35%), el metabolismo (22%), el transporte (12%), la replicación y la modificación del DNA (10%), el plegamiento y la degradación de proteínas (8%), y los procesos celulares (6%).

Una de las características más notables del proteoma humano es que tiene muchas proteínas únicas, en comparación con otros eucariontes; pero tiene relativamente pocos dominios proteicos únicos (porciones de proteínas que tienen una función específica). La mayoría de los dominios proteicos parecen ser comunes al reino animal. Existen muchas arquitecturas proteicas únicas que, sin embargo, son definidas como combinaciones únicas de dominios. La FIGURA 5.16 muestra que la mayor proporción de proteínas únicas son proteínas transmembrana y extracelulares. En las levaduras, la gran mayoría de las arquitecturas pertenecen a las proteínas intracelulares. En las moscas (o gusanos), se encuentran alrededor del doble de arquitecturas intracelulares, pero existe una proporción notablemente más elevada de proteínas transmembrana y extracelulares, como se podría esperar a partir de las funciones adicionales requeridas para las interacciones entre las células de un organismo multicelular. Las adiciones en las arquitecturas requeridas en los vertebrados (seres humanos) es relativamente pequeña, pero nuevamente existe una proporción mayor de arquitecturas extracelulares y de membrana.

Se sabe desde hace tiempo que la diferencia genética entre los seres humanos y los chimpancés (nuestro pariente más cercano) es muy pequeña, con ~99% de identidad entre los genomas. En la actualidad, la secuencia del genoma del chimpancé permite que se investigue el 1% de las diferencias del genoma con más detalle, para ver si se pueden identificar las características responsables de la "humanidad". La comparación muestra 35×10^6 sustituciones de nucleótidos (1,2% de las diferencias de secuencia global), 5×10^6 deleciones o inserciones (que torna ~1,5% de la secuencia de eucromatina en específica de cada especie) y muchos reordenamientos cromosómicos. Las proteínas homólogas suelen ser muy similares; el 29% son idénticas y, en la mayoría de los casos, existe sólo uno o dos aminoácidos dife-

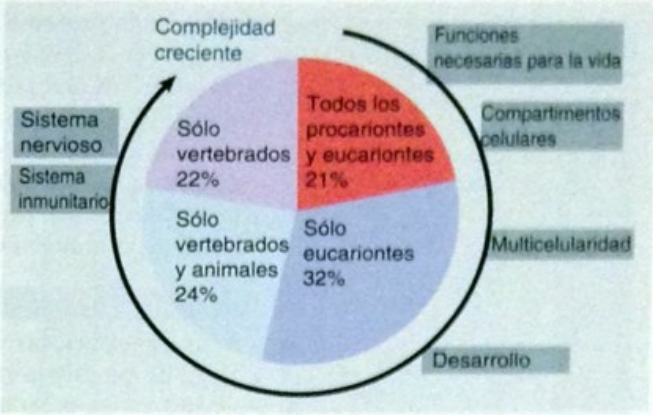


FIGURA 5.14 Los genes humanos se pueden clasificar según cuán ampliamente se distribuyen sus homólogos en las otras especies.

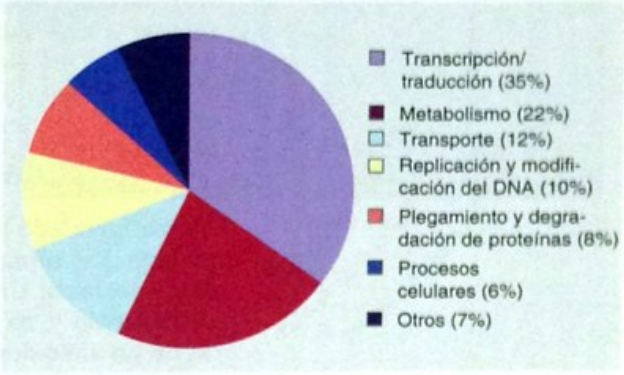


FIGURA 5.15 Las proteínas eucariontes comunes se relacionan con las funciones celulares esenciales.

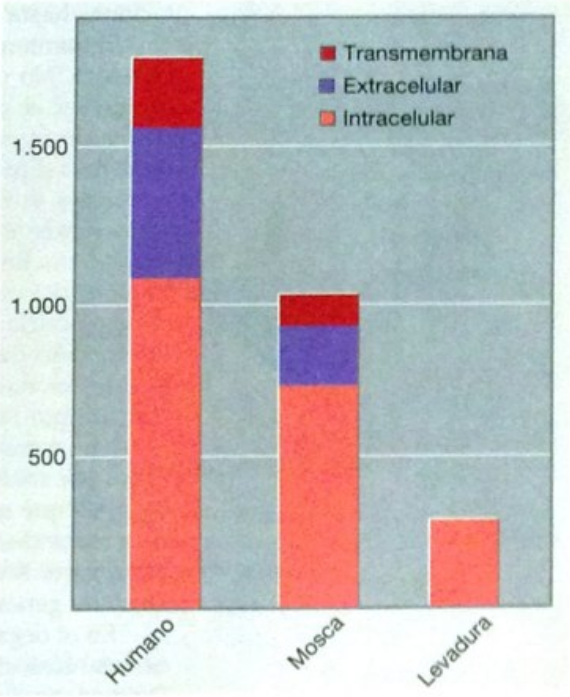


FIGURA 5.16 La complejidad creciente en los eucariontes está acompañada por la acumulación de proteínas nuevas con funciones transmembrana y extracelular.

rentes en la proteína entre las especies. De hecho, las sustituciones de nucleótidos se producen con menos frecuencia en los genes que codifican polipéptidos, ya que es probable que estén involucrados en rasgos específicamente humanos; lo que sugiere que la evolución de proteínas no es un factor importante que explique las diferencias entre el humano y el chimpancé. Esto deja como candidatos principales a los cambios a gran escala en la estructura génica y/o a los cambios en la regulación génica. Un 25% de las sustituciones de nucleótidos se producen en dinucleótidos CpG (entre los cuales están muchos sitios reguladores potenciales).

CONCEPTOS CLAVE

- Las comparaciones de diferentes genomas muestran una correlación positiva entre la cantidad de genes y la complejidad morfológica, en la medida en que se necesitan genes adicionales en los eucariontes, organismos multicelulares, animales y vertebrados.
- La mayoría de los genes que son únicos para los vertebrados se relacionan con los sistemas inmunitario y nervioso.

REVISIÓN DE CONCEPTOS

Explicar la observación de que el proteoma humano tiene muchas proteínas únicas, pero pocos dominios de proteínas únicos.

5.9 ¿Cuántos genes son esenciales?

La fuerza de la selección natural asegura que los genes funcionales se retengan en el genoma. Las mutaciones ocurren al azar, y cualquier efecto sobre un ORF dañará el producto proteico. Un organismo con una mutación dañina estará en una desventaja en la competencia y, en última instancia, la mutación será eliminada. Sin embargo, la frecuencia de un alelo desventajoso en la población se equilibra entre la generación de mutantes nuevos y la eliminación del alelo por selección. Revirtiendo este argumento, siempre que se ve un ORF expresado, intacto en el genoma, se asume que su producto cumple una función útil en el organismo. La selección natural debe evitar que las mutaciones se acumulen en el gen. El destino final de un gen que deja de ser funcional es la de acumular mutaciones hasta que no sea más reconocible como tal.

El mantenimiento de un gen implica que no le confiere una desventaja selectiva al organismo. No obstante, en el curso de la evolución, incluso una pequeña ventaja relativa puede ser el sujeto de la selección natural, y un defecto fenotípico puede no necesariamente ser inmediatamente detectable como el resultado de una mutación. También, en organismos diploides, una mutación recesiva nueva, puede estar "oculta" en la forma de heterocigota durante muchas generaciones. Sin embargo, se debería conocer cuántos genes son realmente esenciales; es decir, aquellos cuya ausencia significaría que es letal para el organismo. En el caso de organismos diploides, implica por supuesto que la mutación nula homocigota es letal.

Se podría asumir que la proporción de genes esenciales disminuirá con el incremento del tamaño del genoma, dado que los genomas más grandes pueden tener múltiples copias relacionadas de funciones génicas particulares. Esta expectativa, aún no ha sido confirmada por la información disponible hasta el momento (véase la Figura 5.2).

Una estrategia permite estimar el número de genes es la de identificar los genes esenciales por análisis mutacional. Si se satura alguna región específica del cromosoma con mutaciones que son letales, las mutaciones se registrarían en el mapa dentro de un número de grupos de complementación que se corresponden con el número de los loci letales en esa región. Mediante la extrapolación al genoma como un todo, se podría calcular el número de genes esenciales totales.

En el organismo con el genoma más pequeño que se conoce (*M. genitalium*), las inserciones aleatorias tienen efectos detectables sólo en alrededor de dos tercios de los genes. De manera similar, un poco menos de la mitad de los genes de *E. coli* parecen ser esenciales. La proporción es incluso menor en la levadura *S. cerevisiae*. Cuando se introdujeron inserciones al azar dentro del genoma en un primer análisis, sólo el 12% fue letal, y otro 14%

tuvo el crecimiento impedido. La mayoría de las inserciones (70%) no tuvo efecto. Una inspección más sistemática sobre la base de la eliminación completa de cada uno de los 5.916 genes (>96% de los genes identificados) demostró que sólo el 18,7% son esenciales para el crecimiento en un medio rico (es decir, cuando se proporcionan completamente los nutrientes). La FIGURA 5.17 muestra que éstos incluyen genes en todas las categorías. La única concentración notable de defectos está en los genes que codifican productos implicados en la síntesis de proteínas, de los cuales ~50% son esenciales. Por supuesto, este enfoque subestima el número de genes que son esenciales para que la levadura viva en estado silvestre, cuando no está bien provista de nutrientes.

La FIGURA 5.18 resume los resultados de un análisis sistemático de los efectos de la pérdida de función génica en el gusano *C. elegans*. Se predijeron las secuencias de genes individuales a partir de la secuencia del genoma y, al dirigir un RNA inhibitorio contra estas secuencias (véase la *Sección 13.8, Los eucariontes contienen RNA reguladores*), se obtuvo una colección grande de gusanos en los cuales se evitó el funcionamiento de un gen (previsto) en cada gusano. Los efectos detectables sobre el fenotipo sólo se observaron en el 10% de estos silenciamientos de genes (*knockout*), lo cual sugiere que la mayoría de éstos no cumple con las funciones esenciales.

Existe una proporción mayor de genes esenciales (21%) entre estos genes de gusanos que tienen sus contrapartes en otros eucariontes, lo que sugiere que los genes ampliamente conservados tienden a tener funciones más básicas. También existe una proporción incrementada de genes esenciales entre aquellos que están presentes en sólo una única copia por genoma haploide, comparado con aquellos donde existen múltiples copias de genes idénticos o relacionados. Esto sugiere que muchos de los genes múltiples podrían ser duplicaciones relativamente recientes, que pueden sustituir funciones una por otra.

El análisis exhaustivo del número de genes esenciales en un eucarionte multicelular se realizó en *Drosophila*, en un intento de correlacionar los aspectos visibles de la estructura cromosómica con el número de unidades génicas funcionales. La noción de esto pudo haberse originado a partir de la presencia de bandas en los cromosomas politénicos de *D. melanogaster*. (Estos cromosomas se encuentran en ciertos estadios del desarrollo y representan una forma física inusualmente extendida, en la que se evidencian una serie de bandas –denominadas formalmente cromómeros–; véase la *Sección 23.7, Los cromosomas politénicos forman bandas que se expanden en el sitio de la expresión génica.*) Desde el momento del concepto inicial de que las bandas podrían representar un orden lineal de los genes, ha existido un intento de correlacionar la organización de los genes con la organización de las bandas. Existen ~5.000 bandas en el conjunto haploide de *D. melanogaster*; las cuales varían en tamaño por encima de un orden de magnitud, pero en promedio existen ~20 kb de DNA por banda.

La estrategia básica se fundamenta en saturar una región cromosómica con mutaciones. En general, las mutaciones se toman como letales, sin analizar la causa de la letalidad. Cualquier mutación que es letal se considera que identifica un locus que es esencial para el organismo. Algunas veces las mutaciones causan efectos deletéreos visibles sin letalidad, en cuyo caso también se las define como loci esenciales. Cuando las mutaciones se colocan en grupos de complementación,

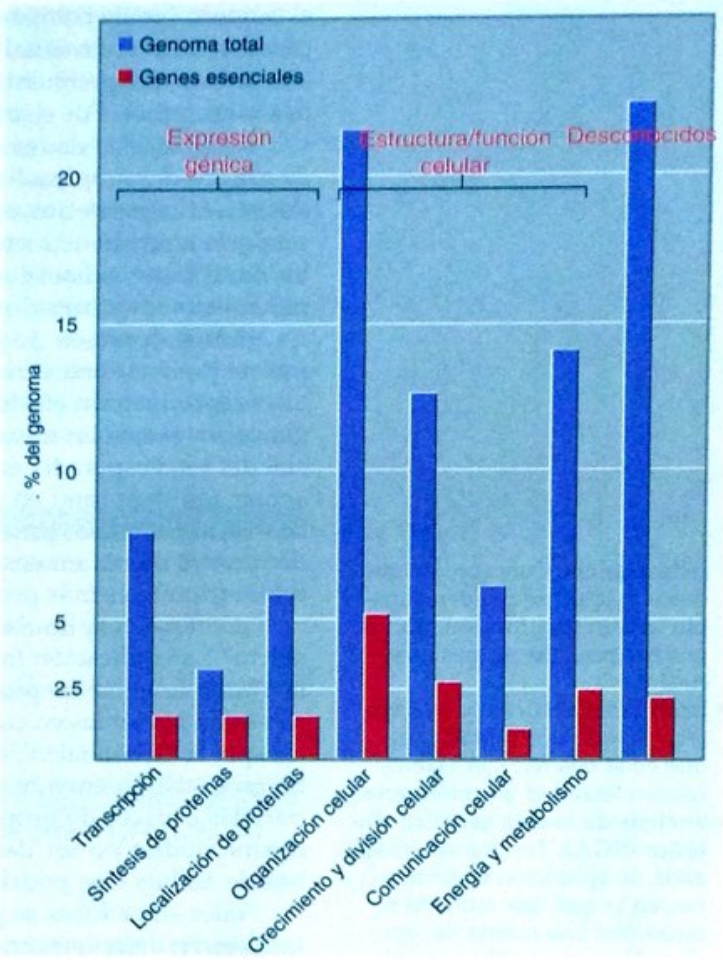


FIGURA 5.17 Los genes esenciales de levadura se encuentran en todas las clases. Las barras azules muestran la proporción total de cada clase de genes, las barras rojas muestran los genes que son esenciales.

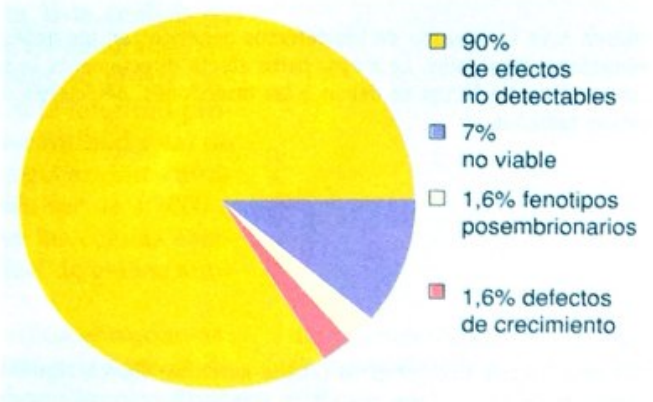


FIGURA 5.18 Un análisis sistemático de la pérdida de función para el 86% de los genes de gusano muestra que sólo un 10% tiene efectos detectables en el fenotipo.

el número puede compararse con el número de bandas en la región, o los grupos de complementación individual pueden incluso ser asignados a bandas individuales. El propósito de estos experimentos es determinar si existe una relación consistente entre las bandas y los genes. Por ejemplo, ¿cada banda contiene un único gen?

Sobre el total de los análisis que se realizaron durante los últimos 35 años, el número de grupos de complementación esencial es ~70% del número de bandas. Una pregunta aún abierta es si existe un significado funcional de esta relación. Independientemente de la causa, la equivalencia otorga una estimación razonable para el número de genes esenciales de ~3.600. Mediante cualquier medida, el número de loci esenciales en *Drosophila* es significativamente menor que el número total de genes.

Si la proporción de genes humanos esenciales es similar a la de otros eucariontes, se podría predecir una variación de ~4.000 a 8.000 genes, en los que las mutaciones serían letales o producirían efectos dañinos evidentes. Hasta el presente, se han identificado 1.300 genes en los que las mutaciones causan defectos evidentes. Esto es una proporción sustancial del total esperado, en especial en vista del hecho de que muchos genes letales pueden actuar tan temprano en el desarrollo que nunca se observarían sus efectos. Esta clase de desviación también puede explicar los resultados que se observan en la **FIGURA 5.19**, que demuestra que la mayoría de los defectos genéticos conocidos se debe a mutaciones puntuales (donde es más probable que se sean, al menos, de alguna función residual del gen).

- **redundancia:** Concepto de que dos o más genes pueden cumplir una misma función, por lo que ninguno de ellos es esencial.
- **letalidad sintética:** La que aparece cuando dos mutaciones que en sí mismas son viables causan letalidad al combinarse.
- **análisis de matriz genética sintética (SGA):** Técnica automatizada de aplicación en levaduras, en la que una mutante se cruza con una matriz de, aproximadamente, 5.000 mutantes con deleciones, para determinar si las mutaciones interactúan para causar un fenotipo sintético letal.

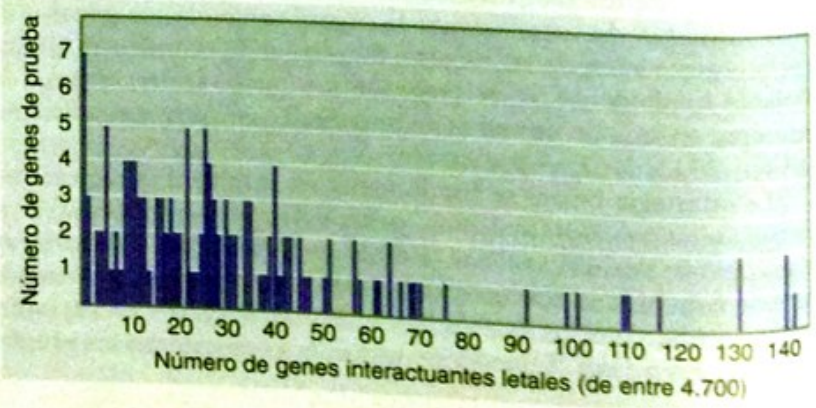
¿Cómo se explica la persistencia de genes cuyas deleciones parecen no tener ningún efecto? La explicación más probable es que el organismo tiene vías alternativas de suplir la misma función. La posibilidad más simple es que exista **redundancia**, y que algunos genes estén presentes en copias múltiples. Esto es verdaderamente cierto en alguno casos, en los que se deben silenciar los genes múltiples (relacionados) para producir un efecto. En un escenario levemente más complejo, un organismo podría tener dos vías bioquímicas separadas capaces de proporcionar alguna actividad. La inactivación de cualquier vía por sí misma podría no ser dañina, pero la ocurrencia simultánea de las mutaciones en los genes de ambas vías podría ser deletérea.

Tales situaciones se pueden evaluar al combinar las mutaciones. En esta estrategia, se introducen deleciones en dos genes, ninguna de las cuales es letal por sí misma, en la misma cepa. Si el doble mutante muere, la cepa se denomina **sintético letal**. Esta técnica se ha utilizado con gran efecto en las levaduras, donde se puede automatizar el aislamiento de los dobles mutantes. Este procedimiento se denomina **análisis de matriz genética sintética (SGA)**. La **FIGURA 5.20** resume los resultados de un análisis en el cual se hizo un SGA para

FIGURA 5.19 La mayoría de los defectos genéticos en los genes humanos se debe a las mutaciones puntuales. La mayor parte afecta directamente la secuencia de proteínas. Los restantes defectos se deben a las inserciones, deleciones o reordenamiento de diversos tamaños.

Mutación sin sentido/mutación terminadora	58%
Corte y empalme	10%
Reguladoras	<1%
Deleciones pequeñas	16%
Inserciones pequeñas	6%
Deleciones grandes	5%
Reordenamientos grandes	2%

FIGURA 5.20 Los 132 genes de prueba mutantes tienen alguna combinación que es letal cuando se combinan con cada una de las 4.700 mutaciones no letales. El gráfico muestra la cantidad de genes de interacción letal existen para cada gen de prueba.



cada una de las 132 deleciones viables, al evaluar si podría sobrevivir en combinación con cualquiera de las 4.700 deleciones viables. Cada uno de los genes evaluados tuvo, al menos, un compañero con el cual la combinación fue letal, y la mayoría de los genes analizados tuvo muchos de tales compañeros; la mediana fue ~25 compañeros, y el mayor número lo muestra un gen analizado que tuvo 146 compañeros letales. Una proporción pequeña (~10%) de los pares de mutantes interactuantes codifica polipéptidos que interactúan físicamente.

Este resultado explicaría la aparente ausencia de efecto de tantas deleciones. La selección natural actuará contra estas deleciones cuando se encuentren en combinaciones letales de a pares. En algún grado, el organismo es protegido contra los efectos del daño de las mutaciones mediante la incorporación de redundancia. Sin embargo, existe un precio en la forma de acumulación de "carga genética" de mutaciones que no son deletéreas en sí mismas, pero que pueden causar problemas graves cuando se combinan con otras mutaciones en generaciones futuras. Presumiblemente, la pérdida de los genes individuales en tales circunstancias produciría una desventaja suficiente como para mantener al gen activo durante el curso de la evolución.

CONCEPTOS CLAVE

- No todos los genes son esenciales. En la levadura y las moscas, la deleción de <50% de los genes tienen efectos detectables.
- Cuando dos o más genes son redundantes, una mutación en cualquiera de ellos puede no tener efectos detectables.
- No se comprende completamente la persistencia de los genes que aparentemente son prescindibles en el genoma.

REVISIÓN DE CONCEPTOS

¿Cuáles son algunos de los problemas en la medición de la cantidad de genes esenciales por análisis mutacional?

5.10 Alrededor de 10.000 genes se expresan en niveles ampliamente diferentes en una célula eucarionte

Se puede determinar la proporción de DNA que contiene genes que codifican proteínas que se expresan en una célula específica en un momento específico mediante la cantidad de DNA que puede hibridarse con los mRNA aislados de esa célula. Este análisis de saturación, realizado para muchos tipos de células en varios momentos, en general, identifica que ~1% del DNA se está expresando como mRNA. A partir de esto, se puede calcular el número de genes que codifican proteínas, siempre que se conozca la longitud promedio de un mRNA. Para un eucarionte unicelular como la levadura, la cantidad total de genes que codifican proteínas es ~4.000. Para los tejidos somáticos de los eucariontes multicelulares, incluyendo tanto plantas como vertebrados, la cantidad suele ser de 10.000 a 15.000. (La única excepción consistente a este tipo de valor la presentan las células cerebrales de mamíferos, en las cuales parecen expresarse una mayor cantidad de genes, aunque el número exacto es incierto).

El número promedio de moléculas de cada mRNA por célula se denomina **abundancia**. Ésta puede calcularse de forma bastante simple si se conoce la masa total de una especie de mRNA específica en la célula. Por ejemplo, en las células del oviducto de pollo, el mRNA total puede representar 100.000 copias de mRNA de ovoalbúmina, 4.000 copias de cada uno de otros 7 u 8 mRNA, y sólo 5 copias de cada uno de los ~13.000 mRNA restantes.

▸ **abundancia:** Número promedio de moléculas de mRNA por célula.

Se puede dividir la población de mRNA en dos clases generales, de acuerdo con su abundancia:

- El oviducto es un caso extremo, con demasiado mRNA representado por una sola especie, pero la mayoría de las células contiene un número pequeño de RNA, cada uno, presente en muchas copias. Este componente de **mRNA abundante** en general consiste en <100 diferentes mRNA presentes en 1.000 a 10.000 copias por célula. A menudo corresponde a la parte principal de la masa; aproximadamente, un 50% del mRNA total.

▸ **mRNA abundante:** Formado por un pequeño número de especies individuales, cada una de ellas presente en una gran cantidad de copias por célula.

- Alrededor de la mitad de la masa del mRNA consiste de un número grande de secuencias, del orden de 10.000, cada una representada por sólo un número pequeño de copias en el mRNA; es decir, <10. Ésta es la clase de **mRNA escaso** o **mRNA complejo.**

mRNA escaso: mRNA formado por un gran número de especies individuales de mRNA, cada una de ellas presente en muy pocas copias por célula. Representa la mayor parte de la complejidad de las secuencias del RNA.

mRNA complejo véase **mRNA escaso**

Muchos tejidos somáticos de eucariontes superiores tienen un número de genes expresados en el orden de 10.000 a 20.000. ¿Cuánto solapamiento existe entre los genes expresados en diferentes tejidos? Por ejemplo, el número de genes expresados en el hígado de pollo es ~11.000 a 17.000, en comparación con el valor para el oviducto que es de ~13.000 a 15.000. ¿Cuántos de estos dos grupos de genes son idénticos? ¿Cuántos son específicos para cada tejido? Existen cuestiones que suelen resolverse mediante análisis de transcriptoma –el conjunto de secuencias que están representadas en el RNA.

Se observa de inmediato que es probable que sean diferencias sustanciales entre los genes expresados en la clase abundante. La ovoalbúmina, por ejemplo, se sintetiza sólo en el oviducto, y no en el hígado. Esto significa que el 50% de la masa del mRNA en el oviducto es específico para ese tejido.

No obstante, los mRNA abundantes representan sólo una proporción pequeña del número de genes expresados. En términos del número total de genes del organismo, y del número de cambios en la transcripción que se deben hacer entre los tipos de células, se necesita conocer la extensión del solapamiento entre los genes representados en las clases de mRNA escaso de los diferentes fenotipos celulares.

Las comparaciones entre los diferentes tejidos muestran que, por ejemplo, ~75% de las secuencias expresadas en el hígado y el oviducto son las mismas. En otras palabras, ~12.000 genes se expresan tanto en el hígado como en el oviducto, ~5.000 genes adicionales se expresan sólo en el hígado y ~3.000 se expresan sólo en el oviducto.

Los mRNA escasos se superponen de gran manera. Entre el hígado y el riñón del ratón, ~90% de los mRNA escasos son idénticos, lo que indica una diferencia entre los tejidos de sólo 1.000 a 2.000 en términos del número de genes expresados. El resultado general obtenido en diversas comparaciones de esta clase es que sólo ~10% de las secuencias de mRNA de una célula son únicas para ésta. La mayoría de las secuencias son comunes a muchos tipos de célula, tal vez incluso a todos.

Esto sugiere que el conjunto común de funciones génicas expresadas, cuya cantidad tal vez sea ~10.000 en los mamíferos, comprenden funciones que son necesarias en todos los tipos celulares. Algunas veces, este tipo de función se denomina **gen de mantenimiento** o **gen constitutivo**. Esta función contrasta con las actividades representadas por las funciones especializadas –tales como la ovoalbúmina o la globina–, necesarias sólo para fenotipos celulares particulares. Éstos suelen denominarse **genes de lujo**.

genes de mantenimiento o constitutivos: Genes que, teóricamente, se expresan en todas las células porque aportan funciones básicas necesarias para el sostén de todos los tipos celulares.

gen de lujo: Gen que codifica una función especializada, expresado en general en grandes cantidades, en tipos celulares determinados.

CONCEPTOS CLAVE

- En una célula determinada, la mayoría de los genes se expresa en niveles bajos.
- Sólo un pequeño número de genes, cuyos productos están especializados para el tipo celular, se expresa altamente.
- Los mRNA expresados en bajos niveles se superponen en gran medida cuando se comparan diferentes tipos celulares.
- Los mRNA expresados en forma abundante suelen ser específicos para el tipo de célula.
- Puede ser común la expresión de ~10.000 genes en la mayoría de los tipos celulares de los eucariontes superiores.

REVISIÓN DE CONCEPTOS

En una célula particular, ¿por qué algunos genes se expresan intensamente y otros escasamente?

5.11 El número de genes expresados puede medirse en grandes cantidades

La tecnología reciente permite estimaciones más sistemáticas y precisas del número de genes expresados que codifican proteínas. Una estrategia (análisis serial de la expresión génica, o SAGE) permite que se utilice una etiqueta de secuencia única para identificar cada mRNA. Así, la tecnología permite medir la abundancia de cada etiqueta. Esta técnica iden-

tifica 4.665 genes expresados en *S. cerevisiae*, siempre que crezcan en condiciones normales, con abundancias que varían desde 0,3 hasta >200 transcriptos/célula. Esto significa que ~75% del número de genes totales (~6.000) se expresa bajo estas condiciones. La FIGURA 5.21 resume el número de mRNA diferentes que se encuentran en cada nivel de abundancia diferente.

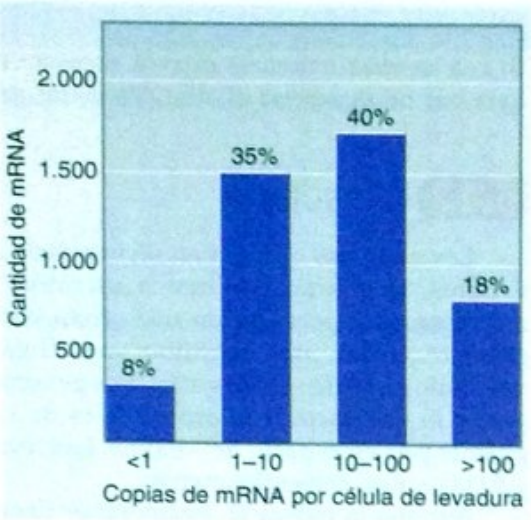


FIGURA 5.21 La abundancia de los mRNA de levadura varía desde <1 por célula (que significa que no siempre cada célula tiene una copia del mRNA) a >100 por célula (que codifican las proteínas más abundantes).

La tecnología más poderosa de la actualidad utiliza *chips* que contienen **micromatrices** –matrices de muchas muestras diminutas de oligonucleótidos de DNA. Su construcción es posible mediante el conocimiento de la secuencia del genoma entero. En el caso de *S. cerevisiae*, cada uno de los 6.181 ORF está representado en la micromatriz por veinte oligonucleótidos de 25 mer, que se aparean perfectamente con la secuencia del mensajero, y con veinte oligonucleótidos con errores de apareamiento que difieren en la posición de una base. El nivel de expresión de cualquier gen se calcula mediante la sustracción de la señal promedio de un error de apareamiento, de su patrón de apareamiento perfecto. El genoma de levadura entero puede ser representado en cuatro *chips*. Esta tecnología es lo suficientemente sensible como para detectar transcriptos de 5.460 genes (~90% del genoma), y demuestra que muchos genes se expresan en niveles bajos, con abundancias de 0,1 a 0,2 transcriptos/célula. Una abundancia de <1 transcripto/célula significa que no todas las células tienen una copia del transcripto en cualquier momento dado.

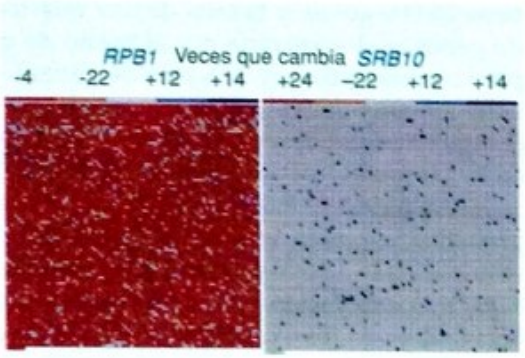


FIGURA 5.22 El análisis de micromatriz de DNA permite medir cambios en la expresión de cada gen. Cada cuadrado representa un gen (la parte superior izquierda es el primer gen en el cromosoma I, la parte inferior derecha es el último gen en el cromosoma XVI). Los cambios en la expresión en relación con el tipo silvestre se indican con el color rojo (reducción), blanco (sin cambio) y azul (incremento). Fotografía cortesía de Rick A. Young, *Whitehead Institute, Massachusetts Insitute of Technology*.

La tecnología permite no sólo mediciones de los niveles de expresión génica, sino también la detección de diferencias en la expresión de células mutantes en comparación con las células de tipo silvestre que proliferan bajo diferentes condiciones de crecimiento, y así sucesivamente. Los resultados de la comparación de dos estados se expresan en una grilla, en la cual cada cuadrado representa un gen particular, y el cambio relativo en la expresión se indica por el color. La parte izquierda de la FIGURA 5.22 muestra el efecto de una mutación en la subunidad más grande de la RNA polimerasa II (*RPB1*), la enzima que sintetiza mRNA, la cual, como se podría esperar, causa la reducción en gran medida de la expresión de la mayoría de los genes. En contraste, la parte derecha muestra que una mutación en un componente auxiliar del aparato de transcripción (*SRB10*) tiene efectos mucho más restrictivos, por lo que causa el incremento en la expresión de algunos genes.

El alcance de esta tecnología a células animales permitirá que las descripciones generales, realizadas sobre la base del análisis de hibridación del RNA, sean reemplazadas por las descripciones exactas de los genes que se expresan, y las abundancias de sus productos, en cualquier tipo celular. Un mapa de expresión génica de *D. melanogaster* detecta la actividad transcripcional en algunos estadios del ciclo vital en casi todos (93%) de los genes pronosticados y muestra que el 40% tiene formas de corte y empalme alternativo.

micromatriz: Serie ordenada de miles de pequeñas muestras de oligonucleótidos de DNA en un pequeño *chip*. Las muestras de mRNA pueden hibridarse con la micromatriz para evaluar la cantidad y el nivel de expresión de los genes.

CONCEPTOS CLAVE

- La tecnología de la micromatriz de DNA permite tomar una foto instantánea de la expresión del genoma entero en una célula de levadura.
- ~75% (~4.500 genes) del genoma de levadura se expresan bajo condiciones de crecimiento normales.
- La tecnología de la micromatriz de DNA permite comparaciones detalladas de las células animales relacionadas, para determinar (por ejemplo) las diferencias en la expresión entre una célula normal y una neoplásica.

REVISIÓN DE CONCEPTOS

Si una levadura unicelular expresa alrededor del 75% de sus genes bajo condiciones normales, ¿por qué no se expresa el otro 25% de los genes?

5.12 Resumen

Los genomas que se han secuenciado incluyen los de muchas bacterias y *archaeas*; levaduras; un gusano, una mosca, un ratón; y el del ser humano. El número mínimo de genes requeridos para formar una célula viviente o parásito es ~470. El número mínimo de genes requerido para constituir una célula de vida libre es ~1.700. Una bacteria gramnegativa típica tiene ~1.500 genes. Los genomas de cepas de *E. coli* varían desde 4.300 a 5.400 genes. El gen bacteriano promedio es de 1.000 pb de longitud y está separado del gen siguiente por un espacio de ~100 pb. Las levaduras *S. pombe* y *S. cerevisiae* tienen entre 5.000 y 6.000 genes, respectivamente.

Aunque la mosca *D. melanogaster* tiene un genoma más grande que el del gusano *C. elegans*, la mosca tiene menos genes (13.600) que el gusano (18.500). La planta *Arabidopsis* tiene 25.000 genes, y la falta de una relación clara entre el tamaño del genoma y el número de genes es demostrada por el hecho de que el genoma del arroz es 4 veces más grande, pero contiene sólo un 50% más de genes (~40.000). Los ratones y los seres humanos tienen entre 20.000 y 25.000 genes, que es mucho menos de lo que se esperaba originalmente. La complejidad del desarrollo de un organismo depende de la naturaleza de las interacciones entre los genes, así como de su cantidad total.

Alrededor de 8.000 genes son comunes en los procariontes y en los eucariontes, y es probable que estén involucrados en funciones básicas. En los organismos multicelulares, se encuentran 12.000 genes adicionales. Otros 8.000 genes se hallan en los animales, y unos 8.000 adicionales (en gran parte involucrados con los sistemas inmunitario y nervioso) se encuentran en los vertebrados. En cada genoma de organismo que ha sido secuenciado, sólo ~50% de los genes tienen funciones definidas. El análisis de los genes letales sugiere que sólo una minoría de los genes son esenciales en cada organismo.

Las secuencias que comprenden un genoma eucarionte pueden clasificarse en tres grupos: las secuencias no repetitivas son únicas; las secuencias moderadamente repetitivas están dispersas y repetidas un número pequeño de veces en la forma de copias relacionadas pero no idénticas; y las secuencias altamente repetitivas son cortas y, en general, se repiten como ordenamientos en tándem. Las proporciones de los tipos de secuencia son características de cada genoma, aunque los genomas más grandes tienden a tener una proporción menor de DNA no repetitivo. Casi el 50% del genoma humano consiste en secuencias repetitivas, la vasta mayoría corresponde a secuencias de transposones. La mayoría de los genes estructurales se localiza en el DNA no repetitivo. La complejidad del DNA no repetitivo refleja mejor la complejidad del organismo, más que la complejidad del genoma total.

Los genes se expresan en una amplia variedad de niveles. Pueden existir 10^5 copias de mRNA para un gen abundante cuya proteína es el principal producto de la célula, 10^3 copias de cada mRNA para <10 mRNA moderadamente abundante, y <10 copias de cada mRNA para >10.000 genes expresados escasamente. Los solapamientos entre las poblaciones de mRNA de las células de diferentes fenotipos son de gran alcance; la mayoría de los mRNA están presentes en muchas células.